C

Calabresi: Heterodox Economic Analysis of Law

Alain Marciano[1,2] and Giovanni Battista Ramello[3,4]
[1]MRE and University of Montpellier, Montpellier, France
[2]Faculté d'Economie, Université de Montpellier and LAMETA-UMR CNRS, Montpellier, France
[3]DiGSPES, University of Eastern Piedmont, Alessandria, Italy
[4]IEL, Torino, Italy

Abstract

Guido Calabresi is one of the founders of the law and economics movement. His approach, however, corresponds to a form of economic analysis of law that, we claim, is heterodox. We show why in this short notice.

Biography

Born on October 18, 1932, in Milan, Guido Calabresi migrated with his family in the USA in 1939. After having received his Bachelor of Science degree (summa cum laude) from Yale College in 1953, majoring in economics, his Bachelor of Arts from Magdalen College at Oxford University in 1955, he got his Bachelor of Laws (LL.B.) magna cum laude from Yale Law School in 1958. Calabresi started by clerking for Justice Hugo Black, who was then US Supreme Court Associate, from 1958 to 1959, and then joined the Yale Law School (instead of the University of Chicago Law School where he was offered full professorship). Calabresi was appointed US Circuit Judge of the US Court of Appeals for the Second Circuit in 1994. He still serves as Sterling Professor Emeritus and Professorial Lecturer in Law at the Yale Law School.

Law and economics emerged just after World War II, gained structure in the 1950s, took further shape in the 1960s, and established itself in the 1970s, essentially under the influence of economists and legal scholars from the so-called Chicago school. The use of economics as a parsimonious tool to tackle otherwise difficult and complex legal problems has now become so frequent that a number of scholars regard this as possibly the most important novelty in all of modern legal scholarship (Denozza 2013; Mattei 1994).

It was Richard Posner who, formally, "invented" the economic analysis of law at the beginning of the 1970s when he published the discipline's eponymous masterpiece (Posner, 1972), launched the *Journal of Legal Studies*, and started to write articles in which he explained that economics is an important tool that can be used to analyze (in particular) legal phenomena. This accounts for the North American, Anglo-Saxon origins of law and economics. But the field was from the outset a melting pot in the broadest sense – not just because it was a field of

A. Marciano, G. B. Ramello (eds.), *Encyclopedia of Law and Economics*,
https://doi.org/10.1007/978-1-4614-7753-2

studies created at the intersection of economics and law but also because it grew out of a blend of North American and European cultures (Ramello 2016). Ronald Coase, another founder of law and economics, was born in England. And Guido Calabresi was born in Italy, where he spent part of his childhood there. His family moved to the USA to escape fascism and brought with them a lively Italian and European bourgeois environment. Even if attempting to deduce impacts from historical backgrounds is generally a risky business (Kalman 2014, p.15), there is strong evidence that Calabresi did indeed blend his family's European culture with that of the USA and that this in turn had an influence on his scholarship. Once, when asked what he considered to be the most important part of his legal education, Calabresi replied:

> "I am a refugee!" And of course, how can I not have been influenced by the fact that we were antifascists and that we left Italy because my father had been jailed and beaten in 1923 and he was a democrat with a small 'd'; that we were very, very rich there and came here with nothing because it was against the law under penalty of death? If I write about capital punishment or if I make a decision, I am not going to be writing to push an agenda but, on the other hand, I would be pretty foolish not to be aware of the fact that that is in my background (Benforado and Hanson 2005, p. 75).

It is not our purpose, here, to look for and find specific traces of European elements in his work, but we nonetheless suggest that Calabresi's work evidences this European heritage. One major aspect of this heritage is what could be called the "comparative" dimension or the adoption of a "comparative viewpoint." This was in particular the case with "Some Thoughts on Risk Distribution and the Law of Torts" (Some Thoughts), Calabresi's first article. Written in the 1950s when Calabresi was still a student, it was published in 1961. That was almost when Coase published his own path-breaking article, "The Problem of Social Cost" (1960). These two works are comparative in the sense that they were using the insights of another discipline to improve another one. While Coase was trying to improve his understanding of economic phenomena by relying on court cases, Calabresi was proposing to use economics to improve one's understanding of economic phenomena. In other words, with "Some Thoughts," Calabresi was probably the first to apply economic methods to analyze legal questions. He really initiated an original approach: of course different from what most legal scholars did but also from what economists – Coase, for instance, or Aaron Director – started to do. Calabresi is not only a founder of the law and economics movement but also of an economic analysis of law.

Commentators on Calabresi's works perceived this innovation at the time of publication. For instance, Walter Blum and Harry Kalven observed the novelty of Calabresi's perspective as soon as they started to study and comment on his work and noted that he had "crystallized the economic analysis of liability" (1967, 240). Posner himself also underscored the change of direction initiated by Calabresi with respect to Coase in his 1970 review of Calabresi's *The Costs of Accidents* (1971). Moreover, in 1971 Frank Michelman, also commenting on *The Costs of Accidents*, noted that Calabresi "provide[s] a conceptual apparatus for describing, comprehending, and evaluating systems of accident law." Strictly speaking, Calabresi's approach was and remains a form of economic analysis of law, because he uses economics as a tool for analyzing legal issues (see Marciano 2012).

Yet Calabresi himself continues to insist that his work should be viewed as a form of law and economics rather than as an economic analysis of law and that he prefers to see his contribution grouped with Coase's rather than with Posner's, whom he strongly disagrees with and even opposes. In his most recent book, he anchored the distinction between his and Posner's approach – to put it differently, between law and economics and an economic analysis of law – in the opposition between Jeremy Bentham and John Stuart Mill. He also insists on the need for an economic approach to law which should rest on a broader cognitive framework than the one used by neoclassical economists. Let us note this very claim was already present in "Some Thoughts," which is indeed remarkable not just because it is foundational for law and economics but also

because it is foundational for Calabresi. In this article, in accordance with the economic analysis of law, Calabresi used economics to guide legal action. But, on the other hand, he recognized that in certain settings "traditional economic theory [can] be of little help," he equally acknowledges the role of laws in fostering economic efficiency, as in the law and economics view. This twofold orientation not only places economics and law on an equal footing, it also treats them both as instruments serving higher goals connected with basic individual liberties, which the market alone is not always able to promote.

For instance, in a 2014 article, Calabresi explained the public function of torts: The liability rule (in both torts and in takings and eminent-domain law) is not used principally, much less solely, to approach the result that would occur in a free market of consensual exchanges (were such a market available) but is instead used approach inalienability (i.e., a fully collective result) in those instances when a criminal law solution is not desired. Calabresi expands on this argument by showing that the liability rule (of the collectively set price) is used to achieve goals that are neither purely libertarian nor purely collectivist but are properly viewed as social democratic.

To understand Calabresi's approach, one must take into account the distinction between choice and consent. Usually, at least in neoclassical economics, individual choices are supposedly made under certain conditions, to which the choosers are assumed to implicitly consent. Consent is thus never discussed or considered in any way distinct from choice. The role of law is precisely to defend consent and to intervene in an efficient way whenever this principle is violated.

Whereas standard economic analyses of law assume that choice means consent, Calabresi insists on the discrepancy between choice and consent, which arises in many practical situations involving legal intervention. To him, the conditions of choice should not be treated as trivially exogenous features of the setting in which legal action – possibly guided by economic efficiency – is played out but rather as a fully fledged part of the decision set, which the legal system must carefully consider. From this viewpoint it follows that the role of law and economics is to provide a method for examining these complex issues and arriving at solutions that consider not only social welfare (and, by implication, efficiency) but also other matters connected to individual rights and liberties (see Marciano and Ramello, 2014).

Thus, Calabresi raised the problem of the potential lack of consent – arising from monopoly, individuals' lack of rationality, and their vulnerability to external pressures, or systemic imperfections – with the implication that individuals do not always make choices that correspond to their preferences. This in its turn enabled law and economics scholars to contribute by providing a wider framework for decision-making that uses the efficiency criterion but also explicitly combines it with other principles, such as societal welfare and individual liberties. The implication is that the questions social scientists have to tackle cannot always be reduced to optimal allocation of resources and instead frequently require enquiring about the "starting points," conditions of choice, and consent to those conditions.

Impact and Legacy

Calabresi is one of the founders of the law and economics movement and was even one of the first to suggest that economics could be used to analyze legal phenomena. One of his main insights and legacy is to have explained that and how law and economics complement each other. To him, economics is concerned with choice under certain given conditions that, as we have noted, may not be satisfactory. What economics provides is only a framework, which needs to be normatively qualified by judges and the legal system. Therefore, if for Calabresi, "law and economics proves the more challenging and worthwhile endeavor" than the economic analysis of law, it is because he envisages law and economics as a back-and-forth dialogue between the two disciplines (Kalman 2014). This equal footing of law and economics is what the economic analysis of law tends to preclude, because it essentially downgrades economics to a mere

problem-solving technology. To be sure, Calabresi sees economics as providing road signs – "road signs that are not too misleading to be worth spending time on" – that judges and lawmakers can then use to serve a higher good than simply fostering efficiency.

Cross-References

- ▶ Coase, Ronald
- ▶ Economic Analysis of Law
- ▶ Posner, Richard

References

Benforado A, Hanson J (2005) The costs of dispositionism: the premature demise of situationist economics. Md Law Rev 64:24–84

Blum W, Kalven H Jr (1967) The empty Cabinet of Dr. Calabresi: auto accidents and general deterrence. Univ Chic Law Rev 34(2):239–273

Calabresi G (1961) Some thoughts on risk distribution and the law of torts. Yale Law J 70(4):499–553

Calabresi G (1971) The costs of accidents: a legal and economic analysis. Yale University Press, New Haven

Calabresi G (2014) A broader view of the cathedral: the significance of the liability rule, correcting a misapprehension. Law Contemp Probl 77(2):1–14

Coase R (1960) The problem of social cost. J Law Econ 3:1–44

Denozza, F. (2013), 'il modello dell'analisi economica del diritto: come si spiega il tanto successo di una tanto debole teoria?', 11(2):Ars Intepretandi, 43–67

Kalman L (2014) Some thoughts on Yale and Guido. Law Contemp Probl 77(2):15–43

Marciano A (2012) Guido Calabresi's economic analysis of law, Coase and the Coase theorem. Int Rev Law Econ 32:110–118

Marciano A, Ramello GB (2014) Consent, choice and Guido Calabresi's Heterodox economic analysis of law. Law Contemp Probl 77(2):97–116

Mattei, U. (1994), 'Efficiency in Legal transplants: an Essay in Comparative Law and Economics', International Review of Law and Economics, 14:13–19

Posner RA (1972) Economic analysis of law. Little, Brown and Co., New York

Michelman F (1971) Pollution as a tort: a non-accidental perspective on Calabresi's costs. Yale Law J 80:647

Ramello GB (2016) The past, present and future of comparative law and economics. In: Ramello G (ed) Comparative law and economics. Edward Elgar Publishing, Cheltenham, pp 3–22

Cameralism

Andre Wakefield
Pitzer College, Claremont Colleges, Claremont, CA, USA

Definition

Cameralism was an aspiring profession during the seventeenth and eighteenth centuries; it thrived in the small territories of the Holy Roman Empire. Academic cameralists, using law and medicine as their models, constructed a system of auxiliary sciences – largely natural, economic, and technological sciences – to support the training of future state servants in the German lands. This system of professional knowledge, known as the cameral sciences, was taught at German universities during the eighteenth century. As a professional model, cameralism ultimately lost out to jurisprudence, but the discourse that it spawned extended well beyond the German lands into Austria-Hungary, Scandinavia, and the Italian states.

Introduction

Historians of economic thought often treat their discipline like physics or chemistry, which is to say, they regard it as a positive science. In this, they follow Milton Friedman. "Economics as a positive science," he famously argued, "is a body of tentatively accepted generalizations about economic phenomena that can be used to predict the consequences of changes in circumstances" (Friedman 1953). To regard economics as a positive science has implications for the way we write its history. Chemistry had its phlogiston; astronomy had its Ptolemaic system; and economics had its preclassical period. Lavoisier, Copernicus, and Adam Smith play the heroes, relegating older theories to the dustbin of history. The narrative of positive science has, for quite some time, motivated dictionaries and encyclopedias like this one. As the great Inglis Palgrave put it, such compendia "show what has actually been written

in former times, and hence will enable the reader to trace the progress of economic thought" (Palgrave 1987).

In a world where Adam Smith and his intellectual progeny play the heroes, it becomes clear what is left for English mercantilists, German cameralists, and other "backward" theorists: they play the foils against which the stories of disciplinary progress get written. H. C. Recktenwald's entry in the *New Palgrave* did just that. "Analytical economics, insights into the laws of the market and the study of the interaction between market and state," he explained, "are relatively unknown in the simple textbooks of the cameralists, which show otherwise sound common sense" (Recktenwald 1987).

Cameralism has been variously defined as a German variant of mercantilism, a university science, a theory of government and society, a baroque science, a political science, an early modern economic theory, and an administrative technology (Roscher 1874; Tribe 1988; Small 1909; Lindenfeld 1997; Schumpeter 1954). Karl Marx just called it a "silly mish-mash of notions inflicted on aspiring bureaucrats" (Marx and Engels 1961–1974). There is some truth in all of it. The dominant narrative, however, has long treated cameralism as a subset of English mercantilism. As Recktenwald put it, cameralism "is the specific version of mercantilism, taught and practised in the German principalities (Kleinstaaten) in the 17th and 18th centuries" (Recktenwald 1987). Nineteenth- and early twentieth-century writers discovered broad lines of agreement between mercantilism and cameralism (Roscher 1874). Certainly, cameralism shared important family resemblances with mercantilism – a commitment to statebuilding, in both political and economic terms, through policies such as import substitution and the industrialization of raw materials (Reinert 2005). Cameralism has also been analyzed as an important early model for and an inspiration for alternative approaches to public finance (Backhaus and Wagner 2004).

But there have been dissonant voices along the way. Writing in 1909, the American sociologist Albion Small suggested that historians of economics had mischaracterized German cameralists. "Cameralism," he argued, "was an administrative technology. It was not an inquiry into the abstract principles of wealth, in the Smithian sense" (Small 1909). Small had a good point because cameralists wrote a lot and most of it did not involve what we would call "economics." Magdelena Humpert's bibliography of cameralist literature included more than 14,000 printed sources (Humpert 1937). Not many of those pages (numbering in their millions) included general discussions about balance of trade or bullionism. Open a cameralist text and you will be more likely to find chapters describing lead smelting, gardening, brewing beer, raising pigs, forestry, and hard-rock mining than, say, general principles of trade. These particulars have most often been ignored in accounts of cameralism as a political or economic theory, but they highlight an important fact: cameralism owed much to the *Kammer*, or fiscal chamber, a specialized collegial body dedicated to administering the sovereign finances (Zielenziger 1914).

More recently, Keith Tribe redefined cameralism as "a university science," placing great weight on the context and practice of university instruction and rejecting the significance of administrative practice for the production of cameralist texts. Instead, Tribe looked to the context of pedagogy and discursive formations. "The two prime influences on these texts," he argued, "were the actual teaching situation and the Wolffian philosophy which informed their style and was itself very largely a product of pedagogic practice" (Tribe 1988). Of paramount importance, then, were the "discursive conditions" under which the texts were produced. Tribe's approach represented a radical departure from earlier scholarship on the subject for he treated cameralism as a self-contained academic discourse, separating the production of the cameral sciences completely from the context of fiscal administration. He also greatly expanded the traditional canon of cameralism by examining hundreds of cameralist texts, many of which were used in university instruction.

Tribe's intervention, in turn, prompted other scholars to think more systematically about the relationship between the cameral sciences and

administrative practice. Perhaps, as Tribe argued, there was no necessary relationship between the two; or perhaps, as Small and others had long maintained, the cameral sciences reflected administrative practice. For some of the more prominent cameralist authors, however, it turns out that there was a relationship between discourse and administrative practice, though not a transparent one. The cameral sciences, that is, did not simply *reflect* everyday practice in the bureaus. Rather, published cameralist texts – canonical works by Veit Ludwig von Seckendorff and Johann von Justi among them – were characterized by a pragmatic utopianism that painted the world as it should be, even as they purported to describe the world as it was. The increasing tendency to treat cameralism as a unique historical formation has challenged the historiographical tradition that relegated the cameral sciences to marginal status in the history of economic thought (Sandl 1999; Tribe 1988; Wakefield 2009).

Historical Development

By the seventeenth century most German territories, large and small, had developed *Kammer* to manage the intimate affairs of princes, dukes, kings, and emperors (Heß 1962; Klinkenborg 1915). By the second half of the seventeenth century, members of the *Kammer* began to be recognized as a distinct group. People started calling them cameralists. Every responsible fiscal official was expected to know his way around a mine or a barley field, because those were the appropriate "ordinary" sources of revenue for his prince, such as income from the mines. ("Extraordinary" sources of revenue, such as direct taxation in times of crisis, were seen as illegitimate and even despotic in many German territories.) Cameralism was structured by the material and institutional realities of fiscal administration in the territories of the Holy Roman Empire.

In the wake of the Thirty Years War, the German lands of the Holy Roman Empire were a mess, devastated, and depopulated. The Peace of Westphalia (1648) recognized more than 300 sovereign territories, ranging widely in size, wealth, and power. For the next 200 years, the Empire served, in the words of Mack Walker, as an "incubator," protecting smaller territories against aggressive incursions from more powerful neighbors (Walker 1971). The economic and political structure of the Empire at once protected and limited the states within it. Cameralists had to accept these limitations, as the ruler of each territory became a kind of entrepreneur seeking to profit from the natural and human resources in his territory.

Insofar as cameralists sought to systematize the daily work of fiscal administration, they faced great obstacles, because the logic of every *Kammer* was distinct, attuned to the local resources of a particular territory or region. The Holy Roman Empire, with its hundreds of kingdoms, duchies, principalities, and bishoprics, presented a staggering diversity of administrative structures, geography, and economic activities. Accordingly, cameralists filled their books with endless detail about the territories in which they lived and worked. This has led authors to suggest that the cameral sciences were descriptive sciences, models of "practical reasoning" that avoided the utopian thinking of nineteenth-century economics (Lindenfeld 1997). It was not, however, always that straightforward. Sometimes, utopian thinking masqueraded as practical, utilitarian knowledge. Cameralists liked to publish "practical" treatises about how to brew beer or raise cattle, for example, and they often made it sound easy. But practical success in agriculture or manufacturing was never easy, which is why failure was the rule when it came to new state ventures. In this respect, cameralists were utopian pragmatists, imagining fields full of healthy crops and fat cows, even as the people drank miserable beer and struggled to feed themselves.

In 1727 Frederick William I of Prussia established the first academic chairs in cameralism, resolving to initiate lectures on *"Cameralia, Oeconomica* and *Polizeisachen*" at his universities in Halle and Frankfurt an der Oder (Stieda 1906; Schrader 1894). "To that end," declared a cabinet order from Berlin, the king had decided to establish a "special Profession, so that students could acquire a good foundation in

these sciences before they are employed in state service." The first "professor of *Cameralia"* was Simon Peter Gasser, a Prussian War and Domains Councilor. Students at the University of Halle were encouraged to attend his lectures, and those who received good recommendations from Gasser could expect special consideration when the time came to appoint new officials. The authorities in Berlin sketched an outline of topics for Gasser's lectures. Frederick William's new "profession" of cameralism demanded sweeping knowledge of Prussia's material circumstances, its productive potential, and the complicated landscape of its rights and privileges.

Cameralism as Profession

After the formal establishment of academic cameralism in 1727, cameralists throughout the Holy Roman Empire sensed an opportunity to establish themselves professionally. We should not imagine these men as members of a political economic school, like the physiocrats, or as some early modern version of the Chicago School of Economics. Cameralist reformers had bigger dreams. They imagined their subject not as a discipline, such as history or mathematics, but as an entirely new academic profession. Cameralism, in other words, would be modeled on law and medicine. Johann von Justi, the most prominent of cameralist proselytizers, was very clear about this in his groundbreaking 1755 cameralist textbook, *Staatswirthschaft*. He suggested that the existing professional faculties of theology, law, and medicine be supplemented by a cameralist faculty (Tribe 1988). The new professors would need to be skilled in areas ranging from forestry and manufactures to taxes and chemistry. "The professor of chemistry would be chosen so that he could lecture on assaying and smelting, and not just the preparation of medicaments. . . the teacher of mechanics would be able to lecture on mining machinery, and the professor of *Naturkunde* would need adequate knowledge about the essence of ores and of deposits." There would be six professors in all, "to which one might add a teacher of civil and military engineering." Not only would this new faculty train skilled future officials, but it would offer "advice for the many institutions and undertakings of the state, for which one must often turn to foreigners at great expense" (Justi 1755a).

Behind the recitation of cameralist principles and material detail in hundreds of textbooks, then, there was a roiling debate about what it meant to be a cameralist. It was a struggle not so much over abstract principles of wealth creation as it was over professional identity. When "aspiring cameralists" flocked to places like Göttingen and Lautern to hear lectures in the cameral sciences, they were not just studying economic policies and the principles of good police, but they were also learning how to behave as members of the *Kammer*. It was not enough to know about budgeting and accounting, one had to be fashionable as well. The classic markers of scholarly culture – knowledge of Latin, learned disputation, reference to authoritative sources, and reading the textbook from a lectern – were rejected in favor of more gentlemanly approaches. Justi, when he arrived in Göttingen, was very specific about this. "I have employed a special teaching style in my courses." He would, contrary to common practice, lecture for only thirty minutes. "Then I got down from my lectern and, standing together with my listeners, I spent the rest of the hour in free and sociable conversation about the lecture material" (Justi 1755b).

It would be a mistake, therefore, to view the cameral sciences as nothing more than a set of political economic principles. For Justi, as for other cameralist reformers, the new profession represented a new way of life and a new epistemology. One could not simply teach students to balance the books and learn about revenue sources; equally important was the need to behave like a proper servant of the Kammer. One had to know how gentlemen acted at court, how to make polite conversation, and how to avoid being a tedious pedant. Cameralism was thus a perfect fit with the new model universities of eighteenth-century Germany, notably Halle and Göttingen (two centers of the cameral sciences). Gerlach Adolf von Münchhausen, Hanoverian minister and first curator of the University of Göttingen,

sought from the very beginning to attract young noblemen and wealthy students to his university. Like the ideal classroom of Justi's reveries, Münchhausen's university aimed to attract fashionable and wealthy students. Münchhausen focused on building nice streets, coffee houses, and impressive academic buildings as a way to attract the right kind of student. For him, utility meant the ability to attract wealthy students to Göttingen from around the Holy Roman Empire, and even from England. For this, one would need famous professors and fashionable knowledge. From this perspective the cameral sciences, fashionable sciences designed to appeal to wealthy noblemen, were perfect. Münchhausen brought Justi to Göttingen in 1755, the same year in which his *Staatswirthschaft,* the most influential of cameralist textbooks, appeared in print (Wakefield 2009).

Cameralism, as imagined by Justi and Münchhausen, was not a stand-alone science like economics or sociology; it was a system of professional education. During the latter half of the eighteenth century, reformers worked to create cameralist faculties throughout the lands of the Holy Roman Empire. Cameralist reformers managed to alter university curricula, establish new academies, and found separate university faculties (Stieda 1906; Tribe 1988; Klippel 1995). In many cases, they even instituted examinations and succeeded in making access to coveted state offices contingent on academic study of the cameral sciences (Bleek 1972). In Göttingen, Münchhausen worked for decades to build a system of "auxiliary sciences" that would create the structure necessary for a cameralist faculty. This involved, most of all, building a system of natural sciences that could serve to train aspiring cameralists. Münchhausen ran into trouble with the other higher faculties – notably law and medicine – in his efforts to harness auxiliary sciences in the service of cameralism. Eventually, though, Münchhausen built a system of sciences, ranging from "economic botany" to "technology," which served as auxiliary sciences to train future state servants (Wakefield 2009).

Göttingen was not alone. In Lautern, 200 miles to the southwest, Friedrich Casimir Medicus founded a freestanding cameralist academy (*Kameral-Hohe-Schule*). Lautern represents the mature example of a professionalizing cameralist curriculum. Hoping to avoid the stubborn traditional faculties and their privileges, Medicus decided to sidestep them altogether by appointing permanent professors to teach subjects such as chemistry, economic botany, technology, and agriculture. For Medicus, it was crucial to have one or two professors dedicated entirely to the natural sciences, because they were the "true foundation upon which all the knowledge of the future state administrator rests, and without which he will never be able to make one sure step forward. One must firmly guarantee that no young man who has failed to study these with zeal is allowed to pass on to the Science of Sources" (Wakefield 2009). Lautern's focus on cultivating the "source sciences" made sense as a strategy for developing the many small and landlocked territories of the Holy Roman Empire. For these principalities and duchies, the constant, intensive improvement of very limited territorial domains – what Sophus Reinert has called "*ersatz* imperialism" – proved more appealing than on the restless, expansionist ambitions of colonial enterprises (Reinert 2011).

Conclusion and Future Directions

The historiography on cameralism stretches back at least two centuries and, as could be expected, that literature records many shifts in approach, definition, and methodology. There is, in other words, no single agreed-upon definition of cameralism; instead, we have a shifting and multifaceted debate about the nature of cameralism and its significance. The most widespread approach to cameralism treats it as a variety of mercantilism, taught and practiced in the German lands of the eighteenth century. Others have treated cameralism as an administrative technology, specifically adapted to the small German territories of the Holy Roman Empire. Still others have defined it as a university science, subject to the pedagogical and discursive conditions present in eighteenth-century German academic settings.

It has also been analyzed as a literature that functioned as public relations for the early modern fiscal policy states of central Europe.

The lack of general agreement about what, exactly, cameralism was (or was not) provides fertile ground for further research. There is much to be done. Recent work has tended to emphasize that cameralist discourse was not limited to the German lands of the Holy Roman Empire, the traditional focus of analysis. Rather, studies in the circulation and translation of texts have revealed that cameralism reached far beyond the German lands, and it was cameralist discourse, not English mercantilism or Smithian political economy, that enjoyed the widest circulation across large swaths of central Europe, Scandinavia, and Italy (Reinert 2011; Lluch 1997). Another burgeoning area of research connects cameralism to technology and the natural sciences. Long seen as peripheral to the "essence" of cameralist discourse, the natural sciences – especially Linnean natural history, chemistry, and mining sciences – have gained increasing attention as core parts of the cameralist enterprise (Koerner 1999, Smith 1994, Wakefield 2000).

References

Backhaus J, Wagner R (2004) Society, state, and public finance: setting the analytical stage. In: Backhaus J, Wagner R (eds) Handbook of public finance. Kluwer, Boston

Bleek W (1972) Von der Kameralausbildung zum Juristenprivileg. Colloquium Verlag, Berlin

Friedman F (1953) Essays in positive economics. University of Chicago Press, Chicago

Heß U (1962) Geheimer Rat und Kabinett in den ernestinischen Staaten Thüringens. Weimar, Böhlau

Humpert M (1937) Bibliographie der Kameralwissenschaften. Köln, Schroeder

Justi JHG (1755a) Staatswirthschaft, 2 vols. Breitkopf, Leipzig

Justi JHG (1755b) Abhandlung von den Mitteln die Erkenntniß in den Oeconomischen und Cameral-Wissenschaften dem gemeinen Wesen recht nützlich zu machen, Göttingen

Klinkenborg M (1915) Die kurfürstliche Kammer und die Begründung des Geheimen Rats in Brandenburg. Hist Z 114(1915):437–488

Klippel D (1995) Johann August Schlettwein and the economic faculty at the University of Gießen. In: Bernard Delmas B, Demals T, Steiner P (eds) La diffusion internationale de la Physiocratie, XVIIIe-XIXe. Grenoble: Presses Universitaires, 1995, pp. 345–365

Koerner L (1999) Linnaeus: nature and nation. Harvard University Press, Cambridge, MA

Lindenfeld D (1997) The practical imagination: the German sciences of state in the nineteenth century. University of Chicago Press, Chicago

Lluch E (1997) Cameralism beyond the germanic world. Hist Econ Id 5:85–99

Marx K, Engels F (1961–1974). Karl Marx, Friedrich Engels. Werke. 39 vols. Dietz, Berlin

Palgrave RHI (1987) Introduction to the first edition. In: Eatwell J, Milgate M, Newman P (eds) The New Palgrave: a dictionary of economics, 3 vols, vol 1. Macmillan, London, p xi

Recktenwald HC (1987) Cameralism. In: Eatwell J, Milgate M, Newman P (eds) The New Palgrave: a dictionary of economics, 3 vols, vol 1. Macmillan, London, pp 313–314

Reinert E (2005) A brief introduction to Veit Ludwig von Seckendorff (1626–1692). Eur J Law Econ 19:221–230

Reinert S (2011) Translating empire: emulation and the origins of political economy. Harvard University Press, Cambridge

Roscher W (1874) Geschichte der Nationalökonomik in Deutschland. Oldenburg, Munich

Sandl M (1999) Ökonomie des Raumes. Der kameralwissenschaftliche Entwurf der Staatswirthschaft im 18. Jahrhundert. Böhlau, Köln

Schumpeter J (1954) History of economic analysis. Oxford University Press, New York

Schrader W (1894) Geschichte der Friedrichs-Universität zu Halle. 2 vols. Berlin: Dummler

Small A (1909) The cameralists, the pioneers of German social polity. University of Chicago Press, Chicago

Smith P (1994) The business of alchemy: science and culture in the Holy Roman Empire. Princeton: Princeton University Press

Stieda W (1906) Die Nationalökonomie als Universitätswissenschaft. Teubner, Leipzig

Tribe K (1988) Governing economy. The reformation of german economic discourse, 1750–1840. Cambridge University Press, Cambridge

Wakefield A (2000) Police chemistry. Sci Context 13:231–267

Wakefield A (2009) The disordered police state: german cameralism as science and practice. University of Chicago Press, Chicago

Walker M (1971) German home towns: community, state, and general estate 1648–1871. Cornell University Press, Ithaca

Zielenziger K (1914) Die alten deutschen Kameralisten. Fischer, Jena

Further Reading

Dittrich E (1974) Die deutschen und österreichischen Kameralisten. Wissenschaftliche Buchgesellschaft, Darmstadt

Nielsen A (1911) Die Entstehung der deutschen Kameralwissenschaft im 17. Jahrhundert. Fischer, Jena
Rosenberg H (1958) Bureaucracy, aristocracy, and autocracy. The prussian experience 1660–1815. Harvard University Press, Cambridge
Sommer L (1920–25) Die Österreichischen Kameralisten. 2 vols. Konegen, Vienna
Troitzsch T (1966) Ansätze technologischen Denkens bei den Kameralisten des 17. und 18. Jahrhunderts. Duncker & Humblot, Berlin

Cap-and-Trade

► Emissions Trading

Cap-and-Trade Approach

► Transferable Discharge Permits

Capitalism

Silvia Ručinská[1], Ronny Müller[1] and Jannik A. Nauerth[2]
[1]Faculty of Public Administration, Pavol Jozef Šafárik University in Košice, Košice, Slovakia
[2]Faculty of Business and Economics, University of Technology Dresden, Dresden, Germany

Abstract

Capitalism is a social and economical system which applies the use of production factors of the economy as a whole. Capitalism can also be understood as a manufacturing process, which is focused on profit maximization and is based on the principle of the invisible hand. Capitalism is associated with the private production of goods and the individual benefit of surplus. These attributes differ from other systems, especially from socialism or communism. Differences occur in the explanation of capitalism, depending on the socioeconomic origin of the explanation.

Additionally, a historical phase of society's development is called capitalism.

Introduction

In general, the economic systems could be described according to coordination mechanism – which can be market or plan coordination and according to ownership, which can be state or private ownership. Each society has a set of production factors that – independent from the social and economic order – should be used for an optimal production and allocation of goods. The main task of any economic and social system is to match production and the preferences of its members. Therefore, it is relevant to decide which goods will be produced and how they will be allocated. To manage this allocation and production decisions, many approaches have been tried in history. There were and still are feudalistic systems without industrial production and an autocratic government that rules economy regarding its own advantage. Besides that there were and still are socialistic systems that used central planning of production to solve the allocation production problems. In addition to the feudal, socialist (or communistic) way of using of the production factors, capitalism is another possible system of providing goods and is formed as a combination of private ownership and market coordination mechanism. Capitalism is described by its elements.

Elements of Capitalism

Capitalism can be described by the following typical elements: private property, freedom, free markets and free competition, private capital and profits, and free prices and wages (Darcy 1970).

The basic of any capitalistic economy is **private property and ownership**. The production factors are not owned by state, and individuals should decide, with respect to their benefit, whether to demand or supply goods. Therefore, individuals have to determine their offered work

force and their accumulation of capital. It is assumed that any individual knows best about their preferences and is able to satisfy them most suitably. Besides that firms produce goods and individuals benefit from the profits. Firms demand labor force and capital to facilitate production and innovations.

In contrast to central planned economies there is no need of a general aggregation of preferences. Therefore, capitalistic economies are based on the sum of decentralized decisions made by individuals. Thus an important characteristic of capitalism is individualism and individual satisfaction of needs.

Other main requirements for capitalistic economies are **free markets** and **free competition**, to achieve efficiency and entrepreneurship. Capitalism is based on the functioning of market mechanism, which is built on the principle of the invisible hand. The principle was first introduced by A. Smith, who stated that individualism and meeting individual prospects by every individual lead to efficiency and benefit the whole economy (Smith 2012). Free competition without restriction by access to information, market access, market structure, and interventions of the government leads the firms to invest in new technology, innovation, and new skills (Hodgson 2003). Also prices and wages are determined by the free market. This idea also assumes that markets are well functioning and a significant state intervention isn't needed. Precondition for free market and private ownership is **freedom.** For some authoress, this issue is important enough that they equate capitalism and a market system (Lindblom 1980).

In capitalism, societal and social conditions are mainly influenced by **capital**. This is due to its high mobility and its universal applicability. Moreover, it is possible to accumulate it unlimited in contrary to labor force. Individuals with a negligible amount of capital can only influence production decisions by consuming.

This leads to a main difficulty of capitalism. The market result is highly influenced by initial endowment of an individual. On the one hand this is essential for an individualistic society, on the other hand this can cause unjustified market outcome. Therefore, politics are addressed to avoid societal not accepted market results.

Irrespective of the characteristics above, there is no general definition of capitalism. The explanations of capitalism are highly influenced by ideological background and target group.

Types of Capitalism

First steps of capitalism's development can be found in the period, which is associated with the development of manufacture production, although simple forms of the so-called protocapitalistic trading appeared already in 3300 BC (Mischer 2014). The exact year of its beginning cannot be specified, but we can subdivide the development of capitalism into three stages: early capitalism, industrial capitalism, and late capitalism (Sombart 2001, 2011). **The early capitalism** includes the time from the sixteenth century to the beginning of industrial revolution. This period is characterized by absolutist monarchs and mercantilism. Besides that, France is the world's leading economy. Due to the political absolutism, economy was widely regulated. To avoid another absolutist monarch benefiting from foreign trade, exporting goods was favored over importing. In the absence of foreign trade and a dynamic system, economic growth was negligible. The production was mainly concentrated on agricultural goods, embedded in a rigid and hierarchical feudalism. A well-known theory describing this period by Malthus (2007) connects the absence of economic growth with population growth. The so-called Malthusian trap states that any improvement in technology results in higher population, rather than improving the economic situation. Despite this rigid environment, first small steps were undertaken to improve economy. In this regard, international trade and banking became more important, and besides that, manufactories arose and thereby announced the industrial revolution.

With the beginning of Industrialization, the age of **industrial capitalism** began and the United Kingdom took over the world economic leadership. Determining an exact beginning depends

on point of view; however, the invention of the first spinning machines in the United Kingdom depicted a cornerstone. After that, industrial development spread over continental Europe before attaching America. One can observe multitudes of social and economic changes in this time. Obviously there were technological improvements, for example, the steam engine, industrial use of electricity, and chemical industry. Thus, by replacing agriculture with industrial production, a structural chance in economy occurred. As a consequence, the production factors labor and capital became more important and the factor land became less important. Moreover, people moved from rural areas to cities, in particular towards coastal regions. This was due to a higher productivity of labor in industry production than in agriculture. Selling labor force to a company promised more than working at subsistence farming. In conjunction with the urbanization, a decrease in death and birth rate took place. Beyond this, the institutional background changed to the benefit of the people; especially democratic processes were initiated. Furthermore, public goods as schooling and security were implemented step by step. Moreover, property rights in material and intellectual dimensions were implemented. According to the structure of economies and political systems, many shapes of capitalism took place.

The late capitalism began at the end of the nineteenth century and the USA became the leading economy. Former small firms developed to large companies by taking over competitors and expansions. Thus, the number of monopolies and cartels rose. In addition, the linkages between companies increased in dimensions of business connections and ownership structure. Moreover, large banks were founded and financial markets became more important in corporate finance. Parallel to this, economic crisis took place in recurring periods. International trade was based on the so-called gold standard. This means that a currency rested on a fixed amount of gold. Thus, everyone could convert paper money in gold. Therefore, importing led to an outflow and exporting to an inflow of gold. According to that, the price level was driven by foreign trade. This leads to the so-called impossible trinity in economic theory. It declares that a stable foreign exchange rate, free capital movement, and an independent monetary policy cannot be achieved at the same time. In recent past, there were approaches to ease this impossibility, but they were of limited duration.

At the beginning of the twentieth century, world economy was shocked by the two World Wars and the "Great Depression." As a result of these shocks and uncertainty about the future, capitalism produced market results that were not commonly accepted. In some countries socialist systems arose, and the iron curtain took place. The capitalistic economies changed into more regulated systems with frequent interventions. However, they managed to preserve the main mechanics of capitalism. The so-called New Deal implemented by Roosevelt in the 1930s depicted the beginning of market regulation. The plan's aim was to increase purchasing power and restore confidence in economy and included gains in public expenditure. After the World War II, the economies of Western Europe reconstructed their political system by focusing on welfare. To avoid unemployment, public pension schemes and redistribution of money became part of the systems. Moreover, the Asian countries, especially China, improved industrial production. By remembering the expansive foreign trade in gold standard, a modern adaption was implemented in 1944, the "Bretton-Woods-System." This was intended to raise the efficiency of international trade and thus improve economic welfare. To achieve fixed but flexible foreign exchange rates, the US dollar set as anchor currency that was bound to gold. The other members had to follow the monetary police of the USA to fix the foreign exchange within narrow bounds. But this approach crashed in 1970s followed by regional currency snakes; sometimes this period is referred to as embedded liberalism. With the end of "Bretton-Woods," political policy tended to liberalization of economies. This includes a slow reduction of subsidies, lower barriers in trade, and reducing market regulations especially in financial markets.

Theoretical View on Capitalism

The modern capitalism based on the "classical" economy represented by Adam Smith and his "Invisible Hand" (Smith 2012), which is used to explain that pursuing own interest promotes the society. Every individual works for their own consumer needs. However, other people are benefiting from this, because they can buy products with higher quality and quantity. Thus, society is promoted by people regarding their own interest and morality. To use full potential from trade he declared the necessity of free markets and the absence of monopolies and cartels.

Karl Marx developed an opposing theory of capitalism (Callinicos A 2012). He established the so-called Historical Materialism to consider society and economy. It states that social and economic change was not driven by ideas but material foundation of people. According to this, human development is a consequence of material equipment and the mode of production. It decomposes history in five parts. These are primitive communism, slave society, feudalism, capitalism, and socialism/communism. To climb up this scale, it is necessary that mode of production is no longer appropriate. The working masses that do not participate in wealth overcome the regime and implement a new system with suitable mode of production and governmental system. This will repeat until communism is reached. A capitalistic society is split in two opposing parts. On the one hand there are owners of capital and on the other hand the working people. Capitalists maximize their income and workers have to sell their working power to them. Marx assumed capitalists selling goods at market price, but workers income is at subsistence level anyway. According to that, capitalists receive the surplus value, and in a competitive environment capitalists have to reinvest it in production. But depending on subsistence wages, the workers do not have enough money to buy all products. Therefore, crises arise and clear the market from overproduction. Thus, Marx predicted that workers, oppressed by frequent crises and increasing social inequality, will overcome capitalism.

Max Weber used a sociological approach to discuss capitalism. He considered capitalism and its characteristics in the Western civilizations (Weber 2008). He works out that rational behavior of individuals increases, but furthermore social acting exists. In addition, he describes that capitalism is driven by expectations over future income from trade. For him, a main character of occidental capitalism is Protestantism. He states that especially Lutheranism and Calvinism lead to another work ethic. In this opinion, this ethic leads to a religious foundation of higher work effort and diligence. Max Weber introduced the concept of "political capitalism" (Holcombe 2015).

Joseph Schumpeter 2008 assumed that capitalism is not a permanent order in economics (Schumpeter 2008). He argued that improvements in efficiency and innovation will lead to monopolies and cartels. As a result from this concentration of economic power, he predicted the destruction of the societal order capitalism is based on. Therefore, similar to Marx he assumed capitalism as not permanent.

The British economist of the twentieth century John Maynard Keynes assumed that the demand is the crucial variable in an economy (Keynes 1997). His theory based on the idea that capitalism is not able to maintain itself. Economic crisis would destroy it if the government does not intervene. In conclusion, Keynes promoted a "strong state" and countercyclical interventions supporting demand, to reduce effects of crises.

Austrian economists Friedrich August von Hayek and Ludwig von Mises partially to Keynes's theory advocated market economy. Hayek assumed that any public administration is inefficient and expensive. Therefore, he rejected a government policy based on interventions and refused any subsidies to firms or demand. Hayek promoted a "slim state" without interventions, in opposite to Keynes.

Keynes and Hayek were proponents of capitalism and both refused any socialist or feudal economic system. Nowadays, modified versions of their theories are frequently used in economic discussions.

In addition to the Keynesian's ideas, the monetary school of thought with Milton Friedman was established in the twentieth century. In his opinion, government should provide property rights, competition, monetary constitution, and support for underage and depended people (Friedman 2002). Opposing to Keynes he states a natural rate of unemployment based on structural frictions in labor markets. Besides that, he states a narrow connection between the quantity of money and inflation. Thus, an appropriate central banking could dispose the problems of inflation and deflation (Friedman 2005) and stabilize economy. To reduce unemployment in the long run he recommended reforms that reduce those structural frictions (Friedman 1968). Friedman suggested a "slim state" that drives markets only by monetary policy. Especially the idea of widening the monetary supply in financial crisis goes back to him.

In the past and also nowadays there are discussions if capitalism as a system could be stabile. The reason of capitalism's crisis is its very nature. Orientation on profit and surplus value means a concentration of money in the hands of a small group of people and also a dissatisfaction of a big group of people, what has to lead to riots and change of the system.

Conclusion/Future Perspective

The undisputed advantages of capitalistic economy are its high productivity, the efficiency, and its adaptability. In the age of capitalism, a level of wealth has been achieved which had never been reached before.

However, there are major challenges that need to be resolved in the future. There are serious problems in environmental pollution due to globalization (Bhagwati 2007). Moreover, there are difficulties in distribution of wealth (Piketty 2014). Besides that, the structure of welfare states is discussed widely with respect to downsize welfare systems. Additionally, durable unemployment is a problem especially in Europe.

References

Monographs

Bhagwati J (2007) In defense of globalization. Oxford University Press, Oxford
Callinicos A (2012) The revolutionary ideas of Karl Marx. Haymarket Books, Chicago
Darcy RL (1970) Primer on social economics. Colorado State University, Colorado
Friedman M (2002) Capitalism and freedom: fortieth anniversary edition. University of Chicago Press, Chicago
Friedman M (2005) The optimum quantity of money. Aldine Transaction, Piscataway
von Hayek FA (1967) Prices and production. Kelley (Augustus M.) Publishers, New York
Keynes JM (1997) The general theory of employment, interest and money. Prometheus Books, Amherst (New York)
Lindblom C (1980) Jenseits von Markt und Staat. Eine Kritik der politischen und ökonomischen systeme. Klett-Cotta, Stuttgart
Malthus TR (2007) An essay on the principle of population. Dover Publications, Mineola (New York)
Piketty T (2014) Capital in the twenty-first century. Belknap, Cambridge
Schumpeter J (2008) Capitalism, socialism and democracy. Harper Perennial Modern Classics, New York
Smith A (2010) The theory of moral sentiments. Digireads Publishing, New York
Smith A (2012) The wealth of nations. Wordsworth Editions, Hertfordshire
Sombart W (2001) Der moderne Kapitalismus, vol 1. Adegi Graphics LLC, New York
Sombart W (2011) Der moderne Kapitalismus, vol 2. Adegi Graphics LLC, New York
Weber M (2008) The protestant ethic and the spirit of capitalism. Digireads Publishing, New York

Journals

Friedman M (1968) The role of monetary policy. Am Econ Rev 58(1):1–17
Hodgson GM (2003) Capitalism, complexity, and inequality. J Econ Issues XXXVII(2):471–478
Holcombe RG (2015) Political capitalism. Cato J 35(1)
Mischer O (2014) Geschichte einer Wirtschaftsordnung. GeoEpoche der Kapitalismus 69:160–167

Caribbean Piracy

► Piracy, Old Maritime

Cartels and Collusion

Bruce Wardhaugh
School of Law, Queen's University Belfast, Belfast, UK

Abstract

This entry provides an introductory account of cartels and collusion and the means used by European and American law to control such practices. The welfare-reducing and welfare-enhancing features of these cartel and other cartel-type arrangements are discussed to demonstrate the need for considered regulation. Both horizontal and vertical arrangements are analyzed, given their different uses and effects in the economy. However, as some forms of collusive activity are welfare enhancing, these are discussed in an effort to show why regulating such behavior must be done with care. Criminal, administrative, and private sanctions are compared as means of control of such agreements. Other topics briefly discussed the nature of (legally) permitted and prohibited collusive information exchange and noneconomic concerns which may justify collusive behavior.

Introduction

JEL codes: K.21, L.40, L.41, L.42

Adam Smith famously wrote, "People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public or in some contrivance to raise prices" (Smith 1776, Bk I, ch x). By this observation, Smith concisely identifies the essence of collusive activity leading to a cartel, which results in an agreement to raise (or fix) prices and harm the consumer in so doing. However, cartel activity can take forms other than naked price-fixing, these include output restrictions, customer allocation (including bid-rigging agreements), and geographic exclusivity of operations. Whatever practice of this sort the parties choose, the practice can give the parties a degree of monopoly power over their customers (Neils et al. 2011, p. 288) and thus isolate the parties from the competitive rigors of the market. Indeed, the European Court of Justice in *ICI* (at para. 64) has referred to this type of activity as "knowingly substitute[ing] practical cooperation for the risks of competition."

Collusive activities can take place among ostensive competitors at the same level of the supply chain (horizontal arrangements or cartels), at different levels of the supply chain (vertical arrangements or cartels), and among members at the same level of the supply chain with the information necessary to collude passed through a member (typically a distributor or wholesaler) at a different level in the supply chain (these arrangements are known as "hub-and-spoke" or "A-B-C" cartels). Yet not all forms of collusion among competitors are regarded as harmful, indeed certain agreements may be beneficial to consumers. Examples of such activities include: joint ventures (particularly those involving research and development to take advantage of synergies of the participants who each may possess specialized knowledge or access to intellectual property) (Motta 2004, pp. 202–205; Neils et al. 2011, pp. 295–297), cooperative standard setting to ensure that customers can take advantage of network effects (Motta 2004, p. 207), and joint purchasing agreements to take advantage of quantities of scale. But at the same time, if the parties to such an agreement possess sufficient market power, such arrangements can lead to anticompetitive effects (Neils et al. 2011, p. 293).

This entry briefly summarizes the results of some of the vast literature on cartels and other forms of economic collusion. More complete bibliographies of this literature can be found elsewhere (see, e.g., Motta 2004; Neils et al. 2011; Kaplow 2013; Wardhaugh 2014). The entry will first discuss the nature of harm which is commonly ascribed to such economic collusion. The discussion of harm ends by outlining the means by which these sorts of activities are controlled or sanctioned. The entry next turns to a discussion

of the types of cartels (horizontal, vertical, hub and spoke) which one sees in today's marketplace. As the effects of these types of agreements can be different, the varying means by which various legal regimes treat such agreements are considered. The entry next examines the features of industries in which cartelization occurs. The fifth section of the entry briefly discusses collusion and information exchange and explains the economic and legal issues. The final section of the entry considers noneconomic considerations which may be taken into account in the evaluation of collusion.

Harms from Cartels

In a competitive market, goods are sold at the marginal cost of their production. The intersection of the cost-and-demand curves therefore determines the quantity produced and the price of the good. As the analysis of cartels is for all intents and purposes identical to the analysis of monopolies (Faull and Nikpay 2014, pp. 21–22), the insights learned from analysis of monopoly power are readily transferable to the analysis of cartels (Stigler 1964). Indeed it is illuminating to view members of a cartel as "divisions" or "branches" of a single-firm monopolist.

The standard economic analysis of the harm from cartels (and monopolies) views them as problematic for the following five reasons:

1. Cartels appropriate consumer surplus to themselves, at the expense of the consumer (Motta 2004, pp. 41–42).
2. Cartels cause deadweight social loss (Motta 2004, pp. 43–44).
3. The creation and preservation of a cartel involves the waste of valuable social resources (Posner 1975).
4. Cartel activity retards the development of new products and processes, thereby depriving consumers of these possible innovations (Motta 2004, pp. 45–47).
5. Participation in a cartel exacerbates managerial slack or "X-inefficiency" (Leibenstein 1966).

In addition to this economic analysis of cartel harm, there are more normative analyses which identify at least some of the harm caused by these arrangements which occurs when cartelists are not playing by the expected rules of the marketplace and do so in a clandestine manner (Whelan 2007, 2014; MacCulloch 2012; Wardhaugh 2012, 2014). Each of these harms is analyzed below.

Appropriation of Consumer Surplus

Assuming the demand curve for a particular good is downward sloping, if the monopolist (or cartel) reduces production, the price of the good will rise. A monopolist (and hence a cartel) will set its production of goods to reflect the marginal revenue. This latter amount of production is less than the amount that would be produced were the price set at marginal cost. As fewer goods are produced by a cartel, their price will rise. Given that consumer surplus is the difference between the price the consumer paid for a good and the maximum price the consumer would pay for a good (reservation price), the elevation in price by a cartel represents a reduction in consumer surplus. Further, as producer surplus (the difference between price for which a good is sold and the cost to produce the good) is the "opposite side of the coin" of consumer surplus, the consumer's loss is therefore the producer's (cartelist's) gain.

It is this appropriation of consumer surplus which has led some to characterize cartel activity as a form of theft. In this vein, a former European Competition Commissioner has used the term "rip-off" to describe the activities of cartels (Kroes 2009). These words are echoed by Whish (2000, p. 220) who claims, "on both a moral and practical level, there is not a great deal of difference between price-fixing and theft."

Creation of Deadweight Loss

Deadweight loss is frustrated (non-)consumption. In the above case, while the harm done is that the consumer paid "too much" (i.e., a super-competitive price) for the product, the purchaser was able to consume the product, as the cartelized price for the good was less than the consumer's reservation price. If under competitive conditions the price of the good would have been below a

potential consumer's reservation price, but if the cartelized price of the good exceeds the reservation price, the good is not purchased. This frustrated consumption is a social deadweight loss.

Costs of Establishing and Maintaining Cartels

Posner (1975) argues that the realization of and ongoing maintenance of a monopoly position (or cartel) involves the expenditure of resources. In the case of monopolies, this can involve lobbying costs associated with regulation (to keep out competition) and other rent-seeking activities. In the case of cartels, such costs include the costs of keeping the arrangement clandestine and even the costs associated with verifying members' compliance with the terms of the agreement (Marshall and Marx 2012, pp. 130–137). It is not an infrequent practice for cartel members to outsource this "audit function" to a third party perceived as neutral to the participants (Marshall and Marx 2012, pp. 134–135). The resources expended on these cartel-preserving activities are viewed as a form of wasted or nonsocially beneficial expenditure.

Reduced Innovation of Products and Productive Processes

One of the rewards of participation in a cartel is a guaranteed return without the requirement (or effort) of engaging in the competitive process. Following Motta (2004, pp. 45–51) we note that given the agreement among cartelists not to compete, there is no incentive for any cartel member to develop new (i.e., "improved") products and to spend resources improving their products or designing new processes to produce the products more efficiently. Indeed, in contrast to a monopolist (who may be concerned with a potential competitor developing a substitute for the monopolized produce and thus may therefore wish to make some investment lest this sort of unwanted entry occurs), given an agreed "standstill" on development, members of a cartel have even less incentive to invest in new products or productive efficiencies.

Managerial Slack

Managerial slack arises from the agency nature of the owner-manager/employee relationship in a firm (Leibenstein 1966; Jensen and Meckling 1976; Motta 2004, p. 47). In effect while owners (shareholders) care about the return on their investment, they delegate the day-to-day operation of the firm to managers. However, managers (and employees) will maximize their own utility function when carrying out their responsibilities. While they may care about the overall profitability of the firm (particularly when incentivized to do so by their remuneration package), they will also care about other matters, such as the effort which they are required to exert in the performance of their duties. By insulating agents in the firms from the rigors of the competitive process, while at the same time ensuring that sales or profit targets are met, cartel behavior is thus not merely a manifestation of managerial slack, but is also an active response for those who wish to pursue a "quiet life," as John Hicks once remarked.

Normative Concerns

In addition to the economic harms isolated above, the legal literature also locates the harm occasioned by cartel activity in the effect which such conduct has on the competitive process (e.g., Whelan 2007, 2014; MacCulloch 2012; Wardhaugh 2012, 2014). Given that there is an understanding by participants in a market transaction that the exchange occurs under conditions of competition, by participating in a collusive arrangement with its ostensive "competitors," a cartelist violates this expectation, thereby perceived as taking advantage of this position in the marketplace or as bringing into disrepute the fairness of the marketplace by clandestinely not playing by its rules. Recent amendments to the UK's *Enterprise Act 2002* reflect this normative position, as sections 188A and B of that Act (as amended by the *Enterprise and Regulatory Reform Act 2013*) exempt from criminal liability individuals who openly enter into such an agreement or otherwise provide affected customers of the details of such arrangements.

Control of Cartels

Given the pernicious nature of many of these agreements, competition regimes attempt to

control such collaboration. In the EU, Article 101 of the *Treaty for the Functioning of the European Union* (TFEU) regulates agreements among competitors. Paragraph 1 of that Article prohibits:

> . . . all agreements between undertakings, decisions by associations of undertakings and concerted practices which may affect trade between Member States and which have as their object or effect the prevention, restriction or distortion of competition within the internal market, and in particular those which:
> (a) directly or indirectly fix purchase or selling prices or any other trading conditions;
> (b) limit or control production, markets, technical development, or investment;
> (c) share markets or sources of supply;
> (d) apply dissimilar conditions to equivalent transactions with other trading parties, thereby placing them at a competitive disadvantage;
> (e) make the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts.

However, paragraph 3 of the same Article exempts from prohibition agreements which improve the production or distribution of goods or technical progress while ensuring consumers receive a fair share of this benefit. This exemption permits agreements which promote, inter alia, research and development, distribution, technology transfer, and licensing. While the EU bases its appraisal of the legality of a mooted scheme on self-assessment (Regulation 1/2003, recitals 4 and 5) rather than the former regime of prior notification and clearance (Regulation 17/1962), the Commission also publishes Guideline and Block Exemption (e.g., those in EU 2013) to provide prospective collaborators with a safe harbor for their proposed agreement. Such a safe harbor is typically available only if the market shares of the participants to the agreement are below certain thresholds, to ensure that the agreement does not permit its participants to obtain a degree of monopoly power.

In contrast, section 1 of the *Sherman Act* (the relevant US law) prohibits "every contract, combination in the form of a trust or otherwise, otherwise, or conspiracy, in restraint of trade or commerce among the several States, or with foreign nations ...." The US Supreme Court has subsequently interpreted the words of the Act to provide for three classes of agreements, which are given differing degrees of scrutiny, depending on the economic harm which the agreement could potentially cause. The most "pernicious" (*Northern Pacific Railway* at 5) agreements are prohibited per se (as the Supreme Court recognized that such arrangements always tend to restrict competition and reduce output (*CBS* at 19–20)), and violations are proven merely through proof that the agreement fell within the prohibited category. Horizontal price-fixing, horizontal fragmentation/division of the market, and concerted refusals to deal all fall into this category.

At the other end of the scale, "rule of reason" analysis examines the entire commercial and economic context of the agreement to determine its anticompetitive effects (*Continental TV*). If the impugned activity unreasonably restrains competition, then the activity is prohibited, as the *Sherman Act* has been interpreted as prohibiting only "unreasonable" restraints on trade (*Standard Oil* at 57). Vertical resale price maintenance agreements are now analyzed under the rule of reason approach (*Leegin*). The third category, "quick look" analysis, sits between per se illegality and a rule of reason analysis. This intermediate analysis is used when a practice is not regarded as per se illegal, but "an observer with even a rudimentary understanding of economics could conclude that the arrangements in question would have an anticompetitive effect on customers and markets" (*California Dental* at 779). With quick look analysis, once the harm has been identified, the burden of proof shifts on the defendant to show the pro-competitive effects (ibid at 770). The Supreme Court employed this form of analysis in examining television restrictions on US University football games, holding that the ostensive justification (to protect live viewing of the games) did not outweigh the anticompetitive effect of broadcast restrictions (*NCAA*).

Cartel activity is controlled administratively (the EU's means) and/or through the use of criminal sanctions. Canada was the first jurisdiction to introduce criminal penalties for cartel activity in 1889. The USA followed with the *Sherman Act* 1 year later. As of 2013 about 25 jurisdictions

have criminalized some form of this activity (Stephan 2014). Of the EU member states, the UK, Ireland, Estonia, Greece, and Germany (for bid rigging) have some form of criminal penalty (ibid). Administrative sanctions are also employed as an ex ante deterrent. In 2013, the EU Competition Commission meted out fines in the amount of €1 882 975 000 for cartel activity (EU 2014). In addition to using criminal sanctions against hard-core cartel activity, the US authorities will also use administrative sanctions as part of their anti-cartel armory.

While the EU's fine totals may appear to be an impressive deterrent, however, in Becker's (1968) sense, they may be suboptimal. There is evidence of recidivism in Europe (Connor and Helmers 2007; Connor 2010). Nevertheless, in the USA, where the mean prison sentence for cartel activity is 25 months (DOJ 2014), there has been no instance of recidivism since July 23, 1999, when the first non-American was imprisoned for antitrust violations (Werden et al. 2011, p. 6). There has been academic discussion of the merits of introducing criminal sanctions for hard-core activity into Europe (see, e.g., Cseres et al. 2006; Beaton-Wells and Ezrachi 2011). However, this is an unlikely prospect, given that cultural attitudes in Europe may not support this use of criminal law and its harsh sanctions (Stephan 2008; Brisimi and Ioannidou 2011).

Civil damages for cartel violations are available both in Europe and in the USA. In the USA, s 15 of the *Clayton Act* permits recovery of triple damages. There is a strong argument that these so-called triple damages do not overcompensate plaintiffs, but merely top up the award of uncompensated losses, such as prejudgment interest, plaintiff's expenses (e.g., experts' reports), and litigation costs (Lande 1993). The presence of a class action regime facilitates the compensation of affected parties, by facilitating the pursuit of smaller claims. In Europe, the pursuit of civil remedies is governed by national law which governs matters including standing, limitation periods, and damages. Opt-out (American-style) class actions are unknown in Europe. The few European jurisdictions which permit class actions (or "collective actions," as they are generally referred to in Europe) typically do so on an opt-in basis. Given the effort (even if small) required to opt into a collective action, classes are typically small, thus leading to the compensation of few affected parties. In one instance in the UK, as a follow-on action to a finding of a Competition Tribunal, only 150 individuals opted into a class to obtain compensation for the overcharge on replica football kits (Which? 2011). This system clearly undercompensates victims of anticompetitive activity. To the extent that private damages supplement the deterrent effect of public enforcement, the lack of opt-out collective actions also likely under-deters anticompetitive conduct.

Types of Cartels

Cartel behavior typically manifests itself as agreements among competitors at the same level of the supply chain (horizontal cartels), as agreements among entities at different levels of the supply chain (vertical arrangements), or as arrangements where information is passed between members at the same level of the supply chain via an intermediary (usually a wholesaler or distributor) at a different level (so-called "A-B-C" or "hub-and-spoke" cartels).

Horizontal Cartels

Typically agreements among competitors who operate at the same level of the supply chain are regarded with the greatest suspicion by regulators, as these sorts of arrangements are most apt to harm consumers' interests through the ability of the participants' collusion to acquire some degree of monopoly power, and use this power to extract monopoly rents from their customers. Typical of such pernicious arrangements are:

- Price-fixing
- Bid rigging
- Restriction of output
- Allocation of geographic territory
- Allocation of customers

Such arrangements, referred to as "hard-core cartel activity" (e.g., OECD 2000, 2002; WTO

2002), are the subject of prohibition of every legal regime which seeks to control anticompetitive conduct. Indeed, in *Verizon* (at 408) Justice Scalia of the US Supreme Court referred to collusion as "the supreme evil of antitrust" on account of the economic harm these practices inflict.

However, not all forms of cooperation at the horizontal level are necessarily consumer welfare reducing. For instance, agreements on standards permit consumer gains from network effects. Likewise, agreements on research and development permit a joint venture between competitors to share the partners' comparative advantages in skill, industrial property, and other resources in a symbiotic manner which could allow for the development of new products (and possibly share the financial risk associated with such a project) (Faull and Nikpay 2014, p. 891). In the EU context, the legality of such agreements is governed by TFEU Article 101(3); and to facilitate self-assessment, the European Commission has promulgated a number of regulations and guidelines on such agreements. These provide a safe harbor for proposed agreements and are predicated on the parties having a low market share in the relevant product markets. The requirement of a low market share ensures the inability of the parties to obtain significant market power and hence extract monopoly rents via their cooperation. In the USA, the Congress has provided research and development statutory exemptions from the Sherman Act (e.g., *National Cooperative Research Act of 1984* and the *National Cooperative Research and Production Act of 1993*). Other forms of agreements will be judicially scrutinized under either the rule of reason or quick look approaches discussed above. Further, to provide additional guidance, the Department of Justice and Federal Trade Commission have published joint guidelines on such arrangements (e.g., DOJ/FTC 1995; FTC/DOJ 2000).

Vertical Cartels/Agreements

These agreements operate at different levels of the distribution scheme; the most common of these sorts of arrangements is resale price maintenance (RPM), namely, a system where the resale price (usually minimum) of a certain good is set by those higher in the chain (typically a manufacturer or distributor). The traditional view of these sorts of agreements is that they either fix prices or (if the set price is a "recommended" price) facilitate price-fixing. As such this practice was formerly viewed as per se illegal in the USA (*Dr Miles*). Since 2007, it is now examined under the rule of reason (*Leegin*). The economic reasoning (primarily starting with Bork 1993, pp. 280–298) in support of relaxing the per se restriction shows not only that there may be consumer welfare-enhancing effects of these products but also that "vertical restraints are not means of creating restriction of output" (Bork 1993, p. 290). These welfare-enhancing effects most prominently include encouraging pre- and post-sale service that would not be provided where retailers compete on price, due to the problem of free riders (Marvel and McCafferty 1984, pp. 347–349; Mathewson and Winter 1998, pp. 74–75; see also Guidelines on Vertical Restraints, para 224).

In Europe, however, the skepticism of the economic benefits of RPM remains. Although the EU's Regulation 330/2010 on Vertical Agreements provides a safe harbor for vertical arrangements among firms with low market shares (Art 2(2)), it specifically excludes RPM practices from this exemption (Art 4(a)). In theory, as with any arrangement, a particular RPM practice could be justified under TFEU Art 101(3), but a close reading on the Commission's guidance on point (Guidelines on Vertical Restraints, paras 223–229) seems to suggest that procompetitive effects are unlikely to be conclusively demonstrated to the satisfaction of the European enforcement authorities.

Other types of vertical arrangements take such forms as market partitioning (i.e., an exclusive grant of sales tights in a particular geographic area to a particular firm) or customer allocation. The EU regime views market partitioning as a threat to market integration, and since the first case, decided by the European Court (*Consten*), European law has attempted to balance integration concerns with efficiency (Jones and Sufrin 2014, p. 790). As Monti (2002, p. 1066) notes the absolutism of both the Chicago School (that geographic restrictions imposed by firms possessing

low market power are efficiency enhancing) and the Commission (that all territorial restrictions are inefficient and thus suspect) is likely misplaced, with a case-by-case analysis to be the most accurate means of assessing efficiencies. Regulation 330/2010 permits customer allocation in cases of small (less than 30%) market share in both seller's and buyer's market (Art 3(1); see Guidelines on Vertical Restraints, para 169). In cases like this the low market share precludes the sellers from extracting monopoly rents and may permit efficiencies when customer-specific investment is appropriate (Guidelines on Vertical Restraints, paras 172–173).

Hub-and-Spoke Cartels

These arrangements have both vertical and horizontal features, where the communication of information to members at the same level (the horizontal element) is made via a member at a different level (the vertical element). The vertical member is typically a wholesaler or distributor. In the UK, recent examples of such cartels include board games (*Argos*), replica football kits (*JJB Sports*), and the dairy industry (*Tesco*). In the USA, the recent e-books case (*Apple*) is an example. As hub-and-spoke cartels only differ from horizontal and vertical cartels by the means in which information is conveyed (via an intermediary at a different level, rather than directly among competitors), the economic analysis of their harms (or efficiencies) is identical to the above cases.

Industries Susceptible to Cartelization

It is an unfortunate fact that no industry appears to be immune from cartelization (see Jones and Sufrin 2014, p. 681, for a non-exhaustive list of European cartels). However, industries which exhibit certain characteristics tend to be more susceptible to collusion. Following Jones and Sufrin (2014, pp. 664–665; see also Veljanovski 2006, pp. 4–6 and Marshall and Marx 2012, pp. 211–237), we note that cartelization is easier (and hence more predominant) in industries in which:

- There are fewer firms.
- Where high entry barriers exist.
- Where cost structures are similar.
- Where the market is transparent.
- Where trade associations or other means of coordination exist.
- Where buyers have little contraveiling purchasing power.
- Where the industry is operating in depressed conditions or below capacity.
- Where the good is an intermediate product.
- Where the good is homogeneous.

These are merely indicia of industries where such activity occurs; they are neither individually nor cumulatively necessary and sufficient conditions for cartelization. Bid rigging in public-sector construction projects may be an example of a form of collusion in an industry which does not exhibit many of the above features.

Collusion

There is a wide spectrum of activity which runs from explicit agreements to coordinate activity on the market to independent – but parallel – conduct in response to changes in market conditions (Kaplow 2013, pp. 21–49). EU law (TFEU Art 101(1)) prohibits "agreements between undertakings, decisions by associations of undertakings and concerted practices" which are anticompetitive. The concept of an agreement has been liberally interpreted by the European General Court, to mean the parties' expression of "a joint intention to conduct themselves on the market in a specific way" (*Bayer*, para 67). The gravamen of the prohibited conduct is the making of an agreement: it need not be implemented. European law does not require that the agreement be formally accepted, tacit acceptance can suffice (*Ford*, *Sandoz*). On the other hand, a concerted practice exists where parties without taking their activity "to a stage where an agreement properly so-called has been concluded, knowingly substitute[s] for the risks of competition practical co-operation between [them]" (*Hüls*, para 158). This is a wider concept than an agreement, and for the arrangement to run

afoul of European law, it must be implemented on the market (*Hüls*, para 165).

The existence and proof of a concerted practice are two distinct legal issues. Although in *ICI* (at para 109), the ECJ held that a concerted practice could safely be inferred as it could "hardly be conceivable that the same action could be taken spontaneously at the same time, on the same national markets and for the same range of products," subsequent decisions have admonished the Commission for being too ready to find such an infringement. In *Ahlström Osakeyhtiö* (at para 71), the ECJ held that parallel conduct "cannot be regarded as furnishing proof of concentration unless concentration constitutes the only plausible explanation for such conduct." In contrast, American law will seek to establish the existence of "plus factors" which are suggestive – or allow for the inference – of collusion (Kovacic et al. 2011; Kaplow 2013, pp. 109–114). Such plus factors include firms' behavior, which appears not to be in their best interests, supernormal returns within an industry, and interfirm transfers of resources (Kovacic et al. 2011, pp. 415, 435–436).

Noneconomic Considerations

TFEU Article 101(3) permits some forms of collusive behavior if certain efficiencies are met and consumers obtain a "fair share" of the benefit of these efficiencies. The term "fair share" is not an economic one and raises normative considerations (See Witt 2012b). In the 1980s and 1990s, this paragraph was used to justify agreements in the *Synthetic Fibre* and *Dutch Brick Industries* (*Stichting Baksteen*) cases to permit an orderly restructuring of these industries. However, since 1999 the Commission has pursued a more "economic approach" to the enforcement of competition law (Witt 2012a, b, pp. 453–455); and as a consequence it is unlikely that such reasoning would be followed today. Nevertheless, it remains an open question as to whether this approach accurately reflects the wording of the European Treaties (Hodge 2012, pp. 104–114; Witt 2012b, p. 471).

Conclusion

The literature on cartels and collusion is voluminous, and space considerations preclude from considering more than an overview of the main issues associated with this sort of activity. What the literature does show, however, is that in certain forms, such activity is economically beneficial, allowing firms to cooperate to produce new or innovative products or processes, thereby permitting consumers to gain. However, these instances of beneficial cooperation are infrequent. The majority of instances of this sort of collusive activity take the form of conspiracies to fix prices or reduce output, to the disadvantage of the consumer. Given the economic harm these sorts of hard-core cartels inflict, control of them is a priority for any competition regime. Yet, there is no consensus on the means by which such collusion is controlled, with both criminal and administrative (both supplemented by ex post private lawsuits) being the chosen means.

References

Books, Articles, Speeches and Other Documents

Beaton-Wells C, Ezrachi A (eds) (2011) Criminalising cartels: critical studies of an international regulatory movement. Hart, Oxford

Becker G (1968) Crime and punishment: an economic approach. J Polit Econ 76:169–217

Bork RH (1993) The antitrust paradox: a policy at war with itself, rev edn. Basic Books, New York

Brisimi V, Ioannidou M (2011) Criminalizing cartels in Greece: a tale of hasty developments and shaky grounds. World Compet 34:157–176

Connor JM (2010) Recidivism revealed: private international cartels 1990–2009. Compet Policy Int 6:101–127

Connor JM, Helmers CG (2007) Statistics on private international cartels. American Antitrust Institute AAI working paper no 07–01. Available at http://ssrn.com/abstract=1103610. Accessed 4 Oct 2014

Cseres KJ, Schinkel MP, Vogelaar FOW (eds) (2006) Criminalization of competition law enforcement: economic and legal implications for the EU member states. Edward Elgar, Cheltenham

Department of Justice (2014) Criminal enforcement fine and jail charts through fiscal year 2013. Available at http://www.justice.gov/atr/public/criminal/264101.html. Accessed 4 Oct 2014

Department of Justice and Federal Trade Commission (1995) Antitrust guidelines for the licensing of

intellectual property. Available at http://www.justice.gov/atr/public/guidelines/0558.htm. Accessed 4 Oct 2014
EU Competition Commission (2013) EU competition law rules applicable to antitrust enforcement. General block exemption regulations and guidelines, vol II. Brussels, EU. Available at http://ec.europa.eu/competition/antitrust/legislation/handbook_vol_1_en.pdf. Accessed 4 Oct 2014
EU Competition Commission (2014) Cartel statistics. Available at http://ec.europa.eu/competition/cartels/statistics/statistics.pdf. Accessed 4 Oct 2014
Faull J, Nikpay A (2014) The EU law of competition, 5th edn. Oxford University Press, Oxford
Federal Trade Commission and Department of Justice (2000) Antitrust guidelines for collaborations among competitors. Available at http://www.ftc.gov/sites/default/files/documents/public_events/joint-venture-hearings-antitrust-guidelines-collaboration-among-competitors/ftcdojguidelines-2.pdf. Accessed 4 Oct 2014
Hodge TC (2012) Compatible or conflicting: the promotion of a high level of employment and the consumer welfare standard under article 101. William and Mary Bus Law Rev 3:59–138
Jensen M, Meckling WH (1976) Theory of the firm: managerial behavior, agency costs and ownership structure. J Financial Econ 34:305–360
Jones A, Sufrin B (2014) EU competition law: text, cases and materials, 5th edn. Oxford University Press, Oxford
Kaplow L (2013) Competition policy and price fixing. Princeton University Press, Princeton/Oxford
Kovacic WE, Marshall RC, Marx LM, White HL (2011) Plus factors and agreement in antitrust law. Michigan Law Rev 110:394–436
Kroes N (2009) Tackling cartels – a never-ending task. Anti-cartel enforcement: criminal and administrative policy – panel session, Brasilia. Available at http://europa.eu/rapid/pressReleasesAction.do?reference=SPEECH/09/454%26format=HTML%26aged=0%26language=EN%26guiLanguage=en. Accessed 4 Oct 2014
Lande RH (1993) Are antitrust 'treble' damages really single damages. Ohio State Law J 54:115–174
Leibenstein H (1966) Allocative efficiency vs. X-inefficiency. Am Econ Rev 56:392–415
MacCulloch A (2012) The cartel offence: defining an appropriate 'moral space'. Eur Compet J 8:73–93
Marshall RC, Marx LM (2012) The economics of collusion: cartels and bidding rings. MIT Press, London/Cambridge, MA
Marvel HP, McCafferty S (1984) Resale price maintenance and quality certification. Rand J Econ 15:346–359
Mathewson F, Winter R (1998) The law and economics of resale price maintenance. Rev Ind Organ 13:57–84
Monti G (2002) Article 81 EC and public policy. Common Mark Law Rev 39:1057–1099
Motta M (2004) Competition policy: theory and practice. Cambridge University Press, Cambridge
Neils G, Jenkins H, Kavanagh J (2011) Economics for competition lawyers. Oxford University Press, Oxford
Organisation for Economic Co-operation and Development (OECD) (2000) Hard core cartels: 2000. OECD, Paris
Organisation for Economic Co-operation and Development (OECD) Directorate for Financial, Fiscal and Enterprise Affairs Competition Committee (2002) Report on the nature and impact of hard core cartels and sanctions against cartels under national competition laws. DAFFE/COMP(2002)7 OECD, Paris. Available at www.oecd.org/dataoecd/16/20/2081831.pdf. Accessed 4 Oct 2014
Posner R (1975) The social cost of monopolies and regulation. J Polit Econ 83:807–827
Smith A (1776) An inquiry into the nature and causes of the wealth of nations. London: printed for W. Strahan; and T. Cadell
Stephan A (2008) Survey of public attitudes to price-fixing and cartel enforcement in Britain. Compet Law Rev 5:123–145
Stephan A (2014) Four key challenges to the successful criminalization of cartel laws. J Antitrust Enforcement 2:333–362
Stigler G (1964) A theory of oligopoly. J Polit Econ 72:44–59
Veljanovski C (2006) The economics of cartels. Finnish Competition Law Year Book. Available at http://ssrn.com/abstract=975612. Accessed 4 Oct 2014
Wardhaugh B (2012) A normative approach to the criminalisation of cartel activity. Legal Stud 32:369–395
Wardhaugh B (2014) Cartels, markets and crime: a normative justification for the criminalisation of economic collusion. Cambridge University Press, Cambridge
Werden GJ, Hammond SD, Barnett BA (2011) Recidivism eliminated: Cartel enforcement in the United States since 1999. Speech before the: Georgetown global antitrust enforcement symposium. Washington, DC. Available at http://www.justice.gov/atr/public/speeches/275388.pdf. Accessed 4 Oct 2014
Whelan P (2007) A principled argument for personal criminal sanctions as punishment under EC cartel law. Compet Law Rev 4:7–40
Whelan P (2014) The criminalization of European cartel enforcement theoretical, legal, and practical challenges. Oxford University Press, Oxford
Which? (A Consumer Advocacy Group) (2011) JJB sports: a case study in collective action. Available at http://www.which.co.uk/documents/pdf/collective-redress-case-study-which-briefing-258401.pdf. Accessed 4 Oct 2014
Whish RP (2000) Recent developments in community competition law 1998/99. Eur Law Rev 25:219–246
Witt AC (2012a) From Airtours to Ryanair: is the more economic approach to EU merger law really about more economics? Common Mark Law Rev 49:217–246
Witt AC (2012b) Public policy goals under EU competition law – now is the time to set the house in order. Eur Compet J 8:443–471
World Trade Organization (WTO) (2002) Working group on the interaction between trade and competition policy, provisions on hardcore cartels: background note by the secretariat. WT/WGTCP/W/191. Geneva, WTO. 20 June 2002

C

Cases

Commission Decisions
Dutch Brick Industry ("Stichting Baksteen"), Commission Decision of 29 Apr 1994 (IV/34.456) [1994] OJ L-131/15
Synthetic Fibres, Commission Decision of 4 July 1984 (IV/30.810) [1984] OJ L-207/17
European Courts
Case 48/69 ICI v Commission [1972] ECR 619
Case 89/85 etc Ahlström Osakeyhtiö et al v Commission ("Wood Pulp II") [1993] ECR I-1307
Case C-199/92 P Hüls AG v Commission [1999] ECR I-4287
Case C-277/87 Sandoz prodotti farmaceutici SpA v Commission [1990] ECR I-45
Case T-41/96 Bayer AG v Commission [2000] ECR II-3383, affirmed Cases C-2 and 2/01 P [2004] ECR I-23
Cases 56 and 58/64 Établissements Consten S.à.R.L. and Grundig-Verkaufs-GmbH v. Commission [1966] ECR 299
Cases 228 and 229/82 Forde Werke AG and Ford of Europe Inc v Commission [1984] ECR 1129
UK
Argos Limited and Littlewoods Limited v Office of Fair Trading [2004] CAT 24 ("Board Games") (Judgment on Liability)
JJB Sports PLC v Office of Fair Trading and Allsports Limited v Office of Fair Trading [2004] CAT 17 ("Replica Football Kits") (Judgment on Liability)
Tesco Stores Ltd, Tesco Holdings Ltd and Tesco Plc v Office of Fair Trading [2012] CAT 31 ("Dairy") (Judgment on Liability)
US
Broadcast Music, Inc v CBS, 441 US 1 (USSC 1979)
California Dental Association v FTC, 526 US 756 (USSC 1999)
Continental TV, Inc v GTE Sylvania Inc, 433 US 36 (USSC 1977)
Dr Miles v John D. Park & Sons Co, 220 US 373 (USSC 1911)
Leegin Creative Leather Products, Inc v PSKS, Inc, 551 US 877 (USSC 2007)
National Collegiate Athletic Association v Board of Regents of University of Oklahoma, 468 US 85 (USSC 1984)
Northern Pacific Railway v United States, 356 US 1 (USSC 1958)
Standard Oil Co v United States, 221 US 1 (USSC 1911)
United States v Apple Inc, 952 F Supp 2d 638, 2013–2 Trade Cases P 78,447 (SDNY July 10, 2013)
Verizon Communications Inc v Law Offices of Curtis V Trinko, LLP, 540 US 398 (USSC 2004)

Statutes and Other Legal Instruments

UK
Enterprise Act 2002
Enterprise and Regulatory Reform Act 2013
European Union
Commission Regulation (EU) No 330/2010 of 20 April 2010 on the application of Article 101(3) of the Treaty on the Functioning of the European Union to categories of vertical agreements and concerted practices [2010] OJ L-102/1
Council Regulation (EC) No. 1/2003 of 16 December 2002 on the implementation of the rules on competition laid down in Articles 81 and 82 of the Treaty ("Regulation 1/2003") [2003] OJ L-1/1
Guidelines on Vertical Restraints [2010] OJ C-130/1
Regulation No 17: First Regulation implementing Articles 85 and 86 of the Treaty ("Regulation 17/62") OJ 013, 21/02/1962 pp 0204–0211
Treaty for the Functioning of the European Union (Consolidated version) [2012] OJ C-326/47
US
Clayton Act, 15 USC §§ 12–27, 29, 52–53
National Cooperative Research Act of 1984 and the National Cooperative Research and Production Act of 1993, codified together at 15 USC §§ 4301–06
Sherman Act, 15 USC §§ 1–7

Cash Demand

Gerhard Graf
Johamer Qutenberg-Universitat Maine, Nieder-Olm, Germany

Abstract

After a definition of cash demand and its peculiarities, cash demand is compared to the more encompassing money demand. Then, the main economic influences on cash demand are explained. A final point exhibits future developments of cash demand which is threatened by electronic inventions enhancing a cashless society.

Synonyms

Currency demand

Definition

Cash demand is a demand for the legacy currency of a country which is influenced by residents and nonresidents of a country.

The Meaning of Cash Demand

Cash demand is the demand for a currency, mostly banknotes. As coins constitute only a relatively small part of total cash demand, they are usually neglected in the analyses of overall cash demand. The demand for the legacy currency is, in general, influenced by the arguments which also drive macroeconomic money demand in a country. However, cash demand is specific first, because the amount demanded is normally fully accommodated by central banks, so demand growth or changes in the composition of banknotes according to their denomination are provided without macroeconomic or budget constraint consequences (Bartzsch et al. 2011). Second, cash demand underlies special microeconomic considerations which somehow balance in the analysis of total money demand, i.e., demand for cash and deposits. Third, cash demand for currencies of different legacies cannot be assumed to be identical. There are rather decisive distinctions between the demand motives for separate currencies (Fischer et al. 2004).

Cash Demand as a Component of Money Demand

Cash demand is a part of overall money demand, for instance, for M1 or M3. The analysis of total money demand should always take into account that it is not a simple sum of the demand for cash and for deposits since there are often substitution processes going on between the components. Moreover, money demand is confronted with macroeconomic money supply considerations and goals of the monetary policies which are not similarly influential for cash demand alone.

Main Influences on Cash Demand

Cash demand is mainly caused by the transaction motive. Households and businesses need cash in order to accomplish everyday consumption goods transactions. This medium-of-exchange function of cash goes together with consumption expenditures of households within a country. But not all private consumption expenditures are paid in cash. A rising amount of cash transactions is substituted by card payments or other technical possibilities to directly withdraw money from personal deposits. In addition, cash is not used unanimously with all banknote denominations of a currency. There is evidence throughout all currencies that people concentrate on small- and medium-size denominations (Deutsche Bundesbank 2009). Thus, large banknotes are often not demanded for the payment of everyday consumption expenditures, but they serve the second demand motive as store of value. Cash as a store of value has opportunity costs. In quite a lot of countries, people rely on cash even if they could earn interest revenues and even if the opportunity cost is increased by inflationary developments of a currency. But, the use of cash as a store of value eases its future use in payments without relevant transactions costs. Additional national influences on cash demand are exerted by the fact that cash provides anonymity in the transactions which are executed with the help of banknotes. For some observers this peculiarity is decisive for the use of cash in the non-reported or illegal economy. Cash, at least some banknote denominations of a group of economies whose currencies are assumed to be more stable internally and more stable relative to other countries, does not only serve as a store of value within the countries themselves but also as a reliable store of value for people in other countries whose currencies are at the risk of inflation or devaluation. Therefore, cash demand for selected currencies is an international demand. This international demand is not only based on the function of international store of value but, to a certain extent, can also be related to the function of medium of exchange within the respective foreign countries (Deutsche Bundesbank 2012). The use of a currency beyond the national borders is a common phenomenon of the US Dollar, the Euro, and the former DM (European Central Bank 2011).

The Future of Cash Demand

During the last decades, cash is subdued to rising competitive assaults by new electronic

developments which tend to reduce cash payments, even for small transaction values, which is paramount to say that cash will lose its outstanding role as medium of exchange and as store of value. Changing payment habits may lead to a "cashless" society and in the end to a complete breakdown for cash demand (Lippi and Secchi 2009). Up to now, however, there are no signs of a definite abandonment of cash of the most important currencies. So cash demand will be a continuing phenomenon in the future. This tendency is enhanced by a supply side argument. As long as cash in the form of banknotes can carry a favorite symbol of a state and will contribute to government revenues via seigniorage, there will be an enduring cash provision easing further demand developments.

Cross-References

- Central Bank
- Money Laundering
- Shadow Economy
- Tax Evasion by Individuals
- Underground Economy

References

Bartzsch N, Rösl G, Seitz F (2011) Foreign demand for euro banknotes issued in Germany: estimation using direct approaches. Deutsche Bundesbank discussion paper series 1: economic studies No. 20/2011

Deutsche Bundesbank (2009) The development and determinants of euro currency in circulation in Germany. Monthly report. 45–58

Deutsche Bundesbank (2012) The usage, costs and benefits of cash – theory and evidence from macro and micro data. Deutsche Bundesbank, Frankfurt

European Central Bank (2011) The use of euro banknotes – results of two surveys among households and firms. Mo Bull 75–90

Fischer B, Köhler, P, Seitz, F (2004) The demand for euro area currencies: past, present and future. in: European Central Bank working paper series, No. 330 April 2004

Lippi F, Secchi A (2009) Technological change and the households' demand for currency. J Monet Econ 56(2):222–230

Further Reading

Ireland PN (2009) On the welfare cost of inflation and the recent behavior of money demand. Am Econ Rev 99:1040–1052

Walsh CE (2010) Monetary theory and policy, 3rd edn. MIT Press, Cambridge, MA

Causation

Samuel Ferey
CNRS, BETA, University of Lorraine, Nancy, France

Abstract

Causation is often said to be one of the most intricate issues in private law. This entry deals with how causation is considered from a law and economics point of view and contributes to enhance our understanding of causation requirements in the law. First, it deals with the main distinction between *ex ante* and *ex post* approach of causation. This leads to study the relationships between causation requirements and efficiency, and the concept of scope of liability is discussed. Then, this economic approach is extended to more difficult cases implying uncertainty. Last, it provides some insights on how these findings remain relevant when considering bounded rationality.

What Is Causation?

"What then is time? If no one asks me, I know that it is. If I wish to explain it to him who asks, I do not know" said Saint Augustine about time. The same could be said about causation: the common sense intuitively knows what causation is, but neither science nor philosophy provides a unified concept of causation. The law does not escape the difficulty and legal causation is often said to be one of the most confused area in all the private law. As Coleman states: "No course in the first year curriculum is more baffling to the average law student than is torts, and for good reasons. In the first two weeks, the student learns that

causation is necessary for both fault and strict liability. Two weeks later the student learns that causation is meaningless, content-free, a mere buzzword. [. . .] Ordinary lawyers and law professors are as confused about causation and the role it plays in liability and recovery as are their students" (Coleman 1992: 270). This is all the more troublesome that causation requirements are one of the keystones of individual liability and justice: it is often said that it is because the injurer caused harm that he has to compensate the victim and that is why "we need an account of causation that can provide reasons of some weight for imposing liability" (Coleman 1992: 271; Cooter 1987).

In law, many concepts of causation are used – proximate cause, cause in fact, probabilistic causation – but the most common legal causation criterion is the "but for test." The "but for test" has an intuitive content: "A causes B" means that B would not have occurred if A had not happened. The "but for test" relies on *necessity*, but beyond its simplicity, this criterion is challenged by many paradoxes and puzzles like overdetermined causes, concurrent causes, or uncertainty (Hart and Honoré 1985). Courts, legal theory, and legislators are aware of the difficulties (*Restatement (Third) on Torts*) and some new criteria are often discussed in the literature (Wright 1985a; Wright and Puppe 2016; Stapleton 2013).

Law and economics literature challenges these classical views on causation and wonders whether and how causation requirements are needed to optimally design tort law (Ben Shahar 2009). For law and economics scholars, the purpose of causation requirements is not so much to provide reasons for imposing liability, as to provide an optimal design of liability regimes. As Shavell states, "the basic function of causation requirement under strict liability, in others words, is that it furnishes socially appropriate incentives to reduce the risk of harm and to moderate the level of activity by imposing liability equal only to the increase in social costs due to a party's action" (Shavell 2004: 251). From a positive point of view, economics may help to better understand legal doctrines and jurisprudence.

Torts Without Causation? From *Ex Post* to Forward Looking Causation

The paper on the *Problem of the Social Cost* published by Coase in 1960 is often said to be the birth of contemporary law and economics. At first glance, Coase's paper is about externality. Coase criticizes classical pigouvian analysis on externality by demonstrating that, in case of zero transaction costs, efficiency could be reached by private negotiations among parties. The example of a cattle-raiser whose cows destroy crops grown by his neighbor is illuminating. Each unit of cow has a marginal cost (the harm suffered by the farmer measured by the value of the crops destroyed by this cow) and a marginal benefit (the additional profit for the cattle-raiser). First, if transaction costs are zero, people will exchange to reach the efficient level of the externality. The only condition is that people know precisely what the property rights are: if the farmer has the right not to be harmed, the cattle-raiser would buy the right to destroy his crops from him; if the cattle-raiser has the right to destroy crops, the farmer would buy the right from him. Second, if transaction costs are positive, efficiency cannot be restored by mutual advantageous exchanges: the initial property rights delineation and tort law play a crucial role to implement or not efficiency.

But the paper goes much further regarding causation. Coasean reasoning leads to consider that causation is reciprocal: "The question is commonly thought of one in which A inflicts harm on B and what has to be decided is: how should we restrain A? But this is wrong. We are dealing with a problem of reciprocal nature. To avoid the harm to B would inflict harm on A. The real question that has to be decided is: should A be allowed to harm B or should B be allowed to harm A? The problem is to avoid the more serious harm." (Coase 1960: 2). The followers of Coase dug out deeper the idea that causation is reciprocal, and that causation should not play any role at all as soon as the aim of the tort system is to reach efficiency. In other words, causation is not a condition to hold a person liable anymore but the result of an economic reasoning (Landes and

Posner 1983). The same can be said about Calabresi "least cost avoider" principle according to which a party is liable for the harm if he was the person who would have avoided it at the least cost (Calabresi 1970; Calabresi 1975).

This first step has radical consequences: causation and efficiency are redundant inquiries as long as they are considered from an *ex post* perspective (Miceli 1997). But things are different as soon as causation is considered from an *ex ante* perspective. Following Calabresi, law and economics literature considers the injurer's behavior may be viewed as a factor that influences the probability that harm occurs and a cause will be modeled as an increasing in the probability that harm comes about. This probabilistic approach has been criticized (Wright 1985b) but has made it possible to easily integrate causation into economics models. For example, the incidental accident illustrates this feature of causation. In the famous case *Berry v. Borough of Sugar Notch*, 43 A. 240 (1899), an accident occurs due to a falling tree on a streetcar (Shavell 2004: 253–254). An *ex post* causation perspective would consider the high speed of the bus as a but for cause for the accident. But an *ex ante* approach would lead to different conclusions: the probability that a bus be striked by a falling tree does not depend on its speed and would be avoided either the bus had been going slower or faster.

Causation, Scope of Liability, and Magnitude of Damage

Law and economics has added to the literature about causation by clarifying the role that causation requirements – both cause in fact and proximate cause – play in the design of liability regimes. A liability regime is defined by the level of care, the magnitude of compensation, and the causation requirements embedded in the scope of liability. We broadly consider causation issues here and we analyze, first, the role of the scope of liability and, second, the magnitude of the damages to be paid by the injurer due to harm he caused.

The scope of liability has been formally introduced by Shavell at the beginning of the 1980s. Shavell defines the scope of liability as the set of the states of the world where the party is held responsible for harm. Consider the following example: a firm builds a dam and takes a certain level of care below the efficient level. Consider now three possible states of the world: a low flood, a medium flood, and an overwhelming flood. Consequences may be associated to each state of the world: the first one would be contained by the dam and would cause no harm at all (event A), the second one would destroy the dam – because the dam has not been built with sufficient care – and would cause harm H (event B), the third one is so huge that it would destroy the dam even if the firm had chosen the required level of care (event C). In case of unrestricted scope of liability, a party is held responsible even if his behavior could not have prevented the loss (event C). In case of restricted scope of liability, the injurer is not held responsible for some states of the world where he would have been able to eliminate or mitigate the loss (event B). The optimal scope of liability is defined as the set of all the states of the world where the level of care is a necessary cause of the harm. In our case, C is not included in this set insofar as C leads to harm whatever the level of care be. A too restricted scope of liability leads to underdeterrence: some benefits of care are not taken into account by the injurer.

But what are the consequences of an unrestricted scope of liability? In other words, does it matter that the parties be held responsible for losses they did not caused? Suppose a firm pollutes the stream of a river and causes a loss of $100, its benefits are $150 and the cost of a device eliminating pollution is $120. From an efficiency viewpoint, it is efficient for the firm to continue its activity, to pollute and to compensate victims for their losses. The strict liability rule implements efficiency: pollution occurs, victims are compensated, and the device is not bought. But suppose now that the pollution could occur due to external factors for an amount of $100. If the parties are held responsible for the losses they did not cause, the firm is likely to pay up to $200. It would prefer to stop its activity even if it is suboptimal.

In Shavell words, it is useful to distinguish the impact of the scope of liability on the care level on one side and on the activity level on the other side. In our example, a too broad delineation of the scope of liability would not change the level of care (in any case, the firm will not pay for the antipollution device insofar as this device has no effect on the probability that a pollution due to external factor comes out). However, it has a potentially "crushing out" effect on the activity level insofar as the firm would prefer to not produce.

In case of negligence, and supposing that the due level of care is the efficient level, an unrestricted scope of liability does not distort incentives. Even if the risk of liability is large due to this unrestricted scope of liability, the party has the incentive to take the due care level and escapes from any compensation to pay. There is no crushing effect of liability on the activity level for the same reason. However, if the scope of liability becomes too restricted, some of the states of the world where it would have been efficient for the society that precaution be taken, are not taken account by the party. This may lead to inefficient levels of care.

The second issue discussed by law and economics is about the magnitude of damage to be paid by the injurer. A subfield of the law and economics literature on causation has discussed the discontinuity of the individual's cost function at the due care point (Grady 1983, 1989, 2014; Marks 1994; Hylon 2014; Kahan 1989). According to the literature, beyond this point, the injurer pays for the entire damage; below this point, he avoids any liability and pays nothing. Grady proposes to deeper consider causation and counterfactuals. Causation requirement is a counterfactual inquiry insofar as court wonders "what would have happen but for the injurer behavior". But what is the relevant counterfactual? Is it the situation which would have come about without any behavior of the injurer or the situation which would have come about if the injurer had taken the level of due care? In the first case, the losses *caused* is said to be the entire loss L; in the second, the loss *caused* is said to be the incremental part of damage caused by the incremental negligence of the party (compared to the level of care he was required to take). Under such a definition of "caused losses," there is no discontinuity in the cost function anymore which leads to original consequences when the implementation of the care level is uncertain.

In most countries, tort law requires that an injurer cannot compensate the victim for more than the harm he actually caused. The statement seems to be obvious and dates back to the very beginning of the private law. Law and economics scholars have challenged this view. First, such an upper bound may lead to inefficiency. This is the case, for example, when the legal system as a whole badly works: if the probability for an injurer who caused a loss to be caught or suited is less than one, the expected cost of a wrongdoing behavior is decreasing and the level of care is likely to be inefficient. Opening the door to compensations above the loss amounts actually caused would help to reach efficiency. Punitive damages are the most obvious legal mechanisms to reach this objective. Other more complex situations involving several tortfeasors have also been studied. Efficiency arguments could be found in favor of original schemes of decoupled liability to increase the total cost paid by the injurer: the cost to be paid by the injurer is divided in two part, one is equal to the loss and is paid to the victim and the other is a fine to be paid to the State (Miceli 1997). Second, the reverse may occur. Recent researches have focused on the possibility, or not, for an injurer to mitigate the damages paid in case of offsetting benefits. Offsetting benefits come about when the injurer's behavior caused both harms and benefits. Imagine the following example: "While driving his car, the defendant negligently hits the plaintiff, who was on her way to the airport to catch a flight, causing a minor injury to her leg. As a result of the accident, the plaintiff misses her flight, which later crashes" (Porat and Posner 2014: 1168). Should the court take into account the benefits caused by the injurer? Principles that courts could follow in these cases may be found in Porat and Posner (2014).

Causation and Uncertainty

As Shavell states, "in many situations, there is uncertainty about causation. For example, it may be not known which manufacturer out of many sold the product (a drug, a lead paint) that caused the injury, or whether an injury was caused by the defendant or by background factors" (Shavell 2004: 254). Such cases may have consequences on the design of optimal liability rule in terms of efficiency (inadequate or excessive incentives), legal errors, or, more broadly, consistency of the law.

Uncertainty over causation occurs when it is impossible to provide sufficient evidence to prove with certainty that a behavior has actually caused harms. Causation uncertainty has two sides: first, when injurers are unknown with certainty; second, when victims are not perfectly identified. The first case refers to the famous example of alternative causation when two hunters shoot and only one bullet hits the victim. The second case is illustrated in the following example. Imagine a neighborhood where each year a fixed proportion of its inhabitants get sick. Suppose the injurer spreads toxic in the air and therefore increases the risk of this disease. Under such circumstances, it is possible to prove that a proportion of the sick people is due to this negligent behavior but it is impossible to know with certainty which ones are the victims of the wrongdoing behavior and which one would have get sick but for it. The injurer is perfectly known but the victims are not.

Uncertain causation is another name for a lack of information due to insufficient evidence: causation is probabilistic not only *ex ante* but also *ex post*. Different rules may be used to determine whether or not the defendant be held liable. Some are implemented by torts law systems all around the world, some may be found only in certain countries, some are still speculative proposals, studied and discussed by scholars without any legal counterparts. These rules have different interesting properties in terms of efficiency, incentives, corrective justice and legal errors. An economic analysis of these rule leads to characterizing the trade-off between these competing objectives.

First is the rule based on the preponderance of the evidence. A party will be held responsible if it is more probable he caused the harm than not. Usually, the preponderance of the evidence is said to require a probability of 51%. According to Kaye, the first who considers this issue from a legal error perspective, this rule leads to minimizing legal error. Suppose the probability that A caused a $1000 loss is 30%. Regarding the preponderance of the evidence, A should not be held liable. The legal error is low: in 30 cases out of 100, A should have paid $1000 and pays zero and in 70 cases out of 100, A should pay zero and pays zero: the total error is $300. Compare now to a proportional liability rule: A pays the amount of the expected damages caused. The error is now of $700 in 30 cases out of 100 (A pays $300 while he caused a loss of $1000) and $300 in 70 cases out of 100 (A pays $300 while he caused zero). The total error is $420.

Even though this minimizing errors, property depends on how legal error is defined (Ferey and G'Sell 2013), the fact that preponderance of the evidence be a threshold rule leads to unexpected consequences. First, it distorts the *ex ante* incentives and therefore does not lead to efficient deterrence. A party who would be aware that the probability of his behavior is below the threshold has no incentive to take the due level of care and to avoid liability while a party aware that its own probability is above the threshold is over deterred. All or nothing rules play as an unrestricted or a too restricted scope of liability. Inefficient activity levels may be expected. More recently, Levmore (2001) and Porat and Posner (2012) have raised paradoxes due to preponderance of the evidence. Suppose that a party involved in two separate and independent accidents and that the probability that he caused each accident is said to be 0.4. Following the preponderance of evidence, the injurer will be exonerated from any liability in both cases (it is more probable that he has not caused the damage, 0.6). This is the result of the black letter of the preponderance of evidence rule. The puzzle arises by considering that the probability that he caused *at least* one of them is above the threshold ($1–0.6^2 = 0.64$).

The second rule discussed by the law and economics literature is the proportional rule. The compensation paid by a party held responsible is proportional to the probability he had to cause the harm. The market share liability doctrine is one of the most famous examples of a proportional rule. Adopted by the Supreme Court of California (*Sindell v. Abbott Laboratories* 607 P.2d 924 (1980)), this rule is adopted, for example, by the European Group of Tort Law (European Group on Tort Law 2005). Under this rule, each potential injurer compensates the victims up to the probability he had to have caused the harm and pays for the expected damages he caused. Knowing whether such a rule leads to efficiency is still controversial. Some consider that the care level and the activity levels will be efficient under a proportional liability rules, others focus on the free rider problem as such as care is considered as a public good (Rose-Ackerman 1990).

Some have proposed to extend this logic in a more radical way. The risk-based liability regime is an example. Within such a regime, structured on the sole probabilities, an injurer is liable for the risks of losses suffered by the victim. From an *ex ante* perspective, a risk-based approach is compatible with efficient incentives, but many strange effects arise from an *ex post* perspective. Such a regime means that the causal link between the injuring behavior and the losses is broken: to be consistent, such an approach will give the right for a potential victim to get compensation even if he has suffered only the risk to be harmed and not the harm itself. And the actual victims of the loss will be undercompensated and will get back only the risk of harm.

Last, a particularly interesting rule is provided by Porat and Stein. For these authors, uncertainty has to be considered as a decision variable from the parties. If, the court or the victim has no sufficient information *ex post*, it is because the parties did not take sufficient precaution to provide the information. The principle to be applied is therefore the *least evidence providers*: liability has to be bear by the party which could have been the most efficient provider of information. In other words, it is efficient that the ones who could have provided the evidence at the cheapest cost be the ones who should be declared liable (Porat and Stein 2001). Parties have incentives to avoid uncertainty and the efficiency liability regimes may be applied.

C

Causation Beyond Rationality

Recent literature in law and economics has challenged the mechanisms by which law induces people to behave in a certain way. Leading by Kahneman and Tversky, behavioral scientists raised important issues about bounded rationality in general and in the law in particular. Several findings of behavioral law and economics have wide impact and significance on causation. First Kahneman himself elaborates on the cognitive illusion of causation (Kahneman 2011). Biases about competence or about the causation links between a behavior and a result have been extensively studied. An interesting example of cognitive bias over causation is provided by the cases of hypothetical causal links between multiple sclerosis and vaccination against hepatitis B. The temporal succession of the events (vaccination following by the disease) seems sometimes to be sufficient to prove causation and courts seem to be victims of a representative bias (Borghetti 2016: 558).

Second, as probabilistic causation is one of the cornerstone of economic approach, many heuristics and biases leading to poorly assess probabilities could impact how people consider causal relationships. If courts, injured people, and injurers make systematic errors on probabilities, they will be mistaken on the true probabilistic links between an event and harm. The hindsight bias exemplifies how a behavioral approach renews law and economics of causation. The hindsight bias covers the fact that an event could be considered *ex ante* as unlikely and, at the same time, be considered as highly probable once it came about. In others words, people suffer inconsistencies in their assessment of the probability of an event depending on the fact that the event has occurred yet or not. In several papers, Rachlinski tests the hindsight bias in judging to see whether

people tend to overestimate the probability of an event once occurred. (Rachlinski 1998). What is at stake is the very ability of the legal system to implement efficient due care level that would be consistent trough time: "The bias, in general, makes defendants appear more culpable than they really are. The bias can cause judges and juries to find liable even those defendants who attempted to avoid negligence by undertaking all reasonable precautions in foresight. Not only does this seem unjust, but it also might have adverse economic consequences. Any potential defendant who is aware of the implications of the hindsight bias might try to avoid liability by taking an excess of precautions. The hindsight bias thus suggests a problem with the law and economics of negligence" (Rachlinski 1998: 572). A lot has still to be done in this field but notice that law and economics converge to recent works in experimental philosophy questioning causation and intention.

Cross-References

- ▶ Multiple Tortfeasors
- ▶ Nuisance
- ▶ Strict Liability Versus Negligence

References

American Law Institute (2010) Restatement (third) of the law of torts: liability for physical and emotional harm

Ben Shahar O (2009) Causation and foreseeability. In: Faure M (ed) Tort law and economics. Edward Elgar, Cheltenham, pp 83–108

Borghetti JS (2016) Causation in hepatitis B vaccination litigation in France: breaking through scientific uncertainty. Chicago Kent Law Rev 91:543–568

Calabresi G (1970) The costs of accidents. Yale University Press, New Haven

Calabresi G (1975) Concerning cause and the law of torts: an essay for Harry Kalven. Univ Chicago Law Rev 43:69–100

Coase RH (1960) The problem of social cost. J Law Econ 3:1–44

Coleman JL (1992) Risks and wrongs. Cambridge University Press, Cambridge

Cooter RD (1987) Torts as the Union of Liberty and Efficiency: an essay on causation. Chicago Kent Law Rev 63:522–551

European Group on Tort Law (2005) Principles of European tort law. Text and commentary. Springer, Vienna

Ferey S, G'Sell F (2013) Pour une prise en compte des parts de marché dans la détermination de la contribution à la dette de réparation. Recueil Dalloz 41:2709

Grady MF (1983) A new positive theory of negligence. Yale Law J 92:799–829

Grady MF (1989) Untaken precautions. J Leg Stud 18:139–156

Grady MF (2014) Marginal causation and injurer shirking. J Tort Law 7:1–34

Hart HLA, Honoré T (1985) Causation in the law, 2nd edn. Oxford University Press, Oxford

Hylon KN (2014) Information and causation in tort law: generalizing the learned hand test for causation cases. J Tort Law 7:35–64

Kahan M (1989) Causation and incentives to take care under the negligence rule. J Leg Stud 18:427–447

Kahneman D (2011) Thinking, Fast and Slow. Farra, Strauss & Giroux, New York

Kaye D (1982) The limits of the preponderance of the evidence standard: justifiably naked statistical evidence and multiple causation. Am Bar Found Res J 7:487–451

Landes WM, Posner RA (1983) Causation in tort law: an economic approach. J Leg Stud 12:109–134

Levmore S (2001) Conjunction and aggregation. Michigan Law Rev 99:723–756

Marks SV (1994) Discontinuities, causation, and Grady's uncertainty theorem. J Leg Stud 23:287–301

Miceli TJ (1997) Economics of the law. Oxford University Press, Oxford

Porat A, Posner EA (2012) Aggregation and law. Yale Law J 122:2–69

Porat A, Posner EA (2014) Offsetting Benefits. Virginia Law Rev 100:1165–1209

Porat A, Stein A (2001) Tort liability under uncertainty. Oxford University Press, Oxford

Rachlinski JJ (1998) A positive psychological theory of judging in hindsight. Univ Chicago Law Rev 65:571–625

Rose-Ackerman S (1990) Market-share allocations in tort law: strengths and weaknesses. J Leg Stud 19:739–746

Shavell S (2004) Foundations of economic analysis of law. Harvard University Press, Cambridge

Stapleton J (2013) Unnecessary causes. Law Q J 129:39–65

Wright RW (1985a) Causation in tort law. California Law Rev 73:1735–1828

Wright RW (1985b) Actual causation *vs*. probabilistic linkage: the bane of economic analysis. J Leg Stud 14:435–456

Wright RW, Puppe I (2016) Causation: linguistic, philosophical, legal and economic. Chicago Kent Law Rev 91:461–506

Central Bank

Karl-Friedrich Israel
Faculty of Economics and Business Administration, Humboldt-Universität zu Berlin, Berlin, Germany
Department of Law, Economics, and Business, University of Angers, Angers, France

Abstract

Central banks evolved in Europe in the seventeenth and eighteenth centuries as centralized monetary authorities that often served the purpose of financing governments. They were instrumental in the transition from classical commodity-backed currencies to the global fiat money system of the present day. The role and functioning of central banks have changed substantially over time. During the classical gold standard, their main responsibility was to store and exchange gold in correspondence to the currencies in circulation. Today, they play a much more active political role in managing the money supply through various policy instruments. They follow different and sometimes conflicting goals, including price stability, stimulation of economic growth, and cutting unemployment rates. A lively debate has arisen on weighing out different policy goals and the proper role of central banking in the economy. This debate does not lack sharp criticisms, both on purely economical and ethical grounds.

Definition

A central bank is an institution that possesses a legal monopoly over the creation of money. It conducts monetary policies and manages the supply of credit and money in a given territory – a country or a union of countries.

The History, Role, and Critique of Central Banking

Central banks – although often considered to be politically independent – are the monetary authorities of states or unions of states. In contrast to commercial banks, they possess legal monopolies over the issuance of money in given territories. They control and manage the supply of credit and money, which usually exhibits legal tender status, through monetary policy tools, such as open market operations, discount window lending, and changes in reserve requirements. Examples of central banks are the *Federal Reserve System* (FED) in the United States, the *Bank of England*, and the *European Central Bank* (ECB) in the eurozone.

According to Bordo (2007), there are three key goals of modern monetary policy. These are price stability or stability of the value of money; economic stability, that is, smoothing business cycles by offsetting shocks to the economy; and financial stability which essentially means granting credit to commercial banks suffering from liquidity shortages as a *lender of last resort*. Historically, the importance of these goals and the goals as such have changed substantially, along with the monetary regimes within which central banks were acting.

A Brief History of Central Banking

In the following section, a brief historical sketch of the development of central banks as monopolists over the supply of money and credit and the economic debates around the subject is presented.

Historically, moneys have been commodities, most often precious metals such as silver and gold. So it was under commodity money regimes that the first institutions resembling central banks were established. According to Glasner (1998, p.27), banking in general "evolved because it provided two services that proved to be strongly complementary: provision of a medium of exchange and intermediation between borrowers and lenders." Instead of using precious metals directly as coins, people could use money substitutes that represented claims to fixed amounts of precious metals that were stored in the banks' vaults.

> Bank money was less costly than coins for several reasons. Holders of deposits often received interest and bore no losses from the wear and tear of coins.

> They also bore no costs of transporting coins or of protecting them against theft or robbery. And, when making transactions, they could avoid the costs of counting, weighting, and inspecting coins. (Glasner 1998, p. 28)

Obviously, banks were put in a very powerful position. There are several historical examples in which these powers were misused. Italian bankers in England, for example, financed the military expenses of Edward I around 1300, when he faced social upheavals in Wales and Scotland. With the financial support of the Italian bankers, he could gather a much greater army than his predecessors were ever able to (Prestwich 1979).

The temptation for banks to issue more money substitutes than there were precious metals in the vaults was compelling. So the municipal bank of Barcelona, for example, had to suspend convertibility into specie during the Catalonian fight for independence in Spain in the fifteenth century (Usher 1943). In the sixteenth and seventeenth centuries, private but highly regulated banks in Venice financed war expenses by creating deposits against government debt. Due to high government demand for funds, the convertibility had to be suspended and the bank money depreciated against precious metals (de Roover 1976; Lane 1937).

In fact, on the way to complete monopolization of the money supply, military conflicts within and between states played a fundamental role. One of the first institutions that are commonly considered to be central banks is the Swedish *Riksbank* founded in the 1660s. It is also the oldest still existing central bank in the world. Another early example is the *Bank of England* founded in the 1690s. The Swedish authorities tried to prevent interference and misuse of the King, whereas the *Bank of England* was specifically founded with the objective of financing the ongoing conflicts with France. It was only through the establishment of this institution that William III was able to borrow £1,200,000, half of which was used to rebuild the navy. The *Banque de France* is another case in point. It was established after the French Revolution and just before the Napoleonic Wars (1803–1815) in 1800. Malpractices of private banks as well as increasing threats from aggressive foreign countries set strong incentives for the establishment of centralized monetary authorities in other European countries and around the world. All that happened mostly under more or less binding silver, gold, or bimetallic standards, which restricted the powers of central banks as long as convertibility of bank notes into specie and the trust of the people in the currency were not to be jeopardized lightly. However, not before 1816, shortly after the Napoleonic Wars, during which convertibility was suspended, has the British pound for the first time been legally defined as a fixed weight of gold, although Britain had been on a de facto gold standard since 1717, due to an overvaluation of gold relative to silver by Sir Isaac Newton who was at that time Master of the Royal Mint.

In 1844, the *Bank Charter Act* reinforced the gold backing of the pound and marked the beginning of a period known as the classical gold standard that lasted until 1914. Also other currencies such as the franc, the mark (since 1871), and the US dollar were legally defined as fixed amounts of gold. Therefore, the classical gold standard was a period of fixed exchange rates between currencies, which lowered risks and fostered international trade, as has been suggested in various empirical studies (see, e.g., López-Córdova and Meissner 2003 or Flandreau and Maurel 2001).

What is today called the time-inconsistency problem of central banking might have played an important role in this development. On the one hand, the monetary authorities could have exploited their monopoly status on a slow but constant basis by inflating the money supply at a modest rate. On the other hand, this would lower the profit-earning potential under emergencies, such as military invasions or economic crises. In order to effectively exploit the monopoly status in emergencies, it is necessary to build up confidence in the stability and soundness of the currency among the general public during normal times. The authorities have to make a credible commitment to price stability (Kydland and Prescott 1977). The developments toward the universal gold standard can be interpreted as such a commitment (Bordo and Kydland 1995). On the other hand, it might be interpreted as a restriction on

governmental despotism and as a political effect of the spread of classical liberal philosophies and enlightenment ideas that generally endorsed limited state powers. The most important function of central banks in that era was to store and exchange gold reserves in correspondence to the national currencies in circulation.

Although there already existed institutions that resembled central banks in the United States in the nineteenth century, namely, the *First* (1790–1811) and the *Second Bank of the United States* (1816–1836), the *Federal Reserve System* as we know it today was only founded in 1913. Shortly after, with the beginning of the Great War (1914–1918), the era of the classical gold standard ended. The United States which entered the war in 1917 inflated their currency less than the European nations and was thus the only country that de jure remained on a gold standard throughout this period. The German, French, and British currencies depreciated substantially with respect to the US dollar and gold. They returned to gold in the 1920s at an artificially overvalued rate which lead to constant complains about gold or liquidity shortages, particularly in Britain, but also in other European countries (Rothbard 1998). The gold exchange standard that was established during the interwar period was essentially a pound sterling standard, since only the British pound sterling was redeemable in gold. Central banks of other European countries mostly held pounds as reserves. The United States was virtually the only country that remained on the gold standard in the classical sense. During the *Great Depression* in the 1930s, many countries suspended convertibility again. After the catastrophe of the Second World War, with the establishment of the Bretton Woods system in 1945, the British pound lost its status as a reserve currency to the US dollar. The *International Monetary Fund* (IMF) and the *International Bank for Reconstruction and Development* (IBRD), which today is part of the *World Bank*, were founded. The US dollar remained convertible into gold on a fixed rate of $35 per fine ounce. Central banks of other countries were responsible for keeping their currencies in a fixed relationship to the US dollar. Only central banks, but no private citizens, were able to hand in US dollars for redemption at the FED. The system lasted until the fifteenth of August 1971, when President Richard Nixon eventually suspended convertibility of the US dollar in a unilateral act. Since then, the global financial system is based on fiat money, that is, money which derives its value from government law and is independent of any commodity.

During the transition period from the classical gold standard to a fiat money system, several economic and political arguments against the gold standard and in favor of fiat currencies have been brought forward. First of all, it would have facilitated the financing of political projects, such as the *New Deal* under Franklin D. Roosevelt in response to the *Great Depression*. Furthermore, it would not have been necessary to increase interest rates in 1931 to maintain convertibility of the US dollar, after Britain was already forced to suspend convertibility (Romer 2003). A policy of low interest rates can only be maintained as long as needed under fiat currencies. Since the money production is not restricted by the production of any commodity, it renders monetary policy much more flexible. The Keynesian theory of business cycles suggests that economic downturns can be cured by stimulating aggregate demand through expansive monetary policy in combination with deficit spending (Keynes 1936; Krugman 2006). There have, of course, been critics of this view (see, e.g., Hazlitt 1959), but it is today accepted by the majority of economists that central banks should intervene more actively into the economy by adjusting the money supply appropriately. In doing so, they should consider and control, as far as possible, macroeconomic aggregates, such as price inflation, unemployment, and economic growth. Often, objectives of monetary policy are in conflict. Simulating aggregate demand and economic growth through expansive monetary policies, for example, is in conflict to the general goal of price stability. Expanding the money supply can only be done at the risk of higher price inflation. Therefore, conflicting policy goals have to be weighed out. Still today, there is a lively debate on which policy instruments should be applied for which purposes and whether central bank policies should be conducted passively, based on rules, or

whether we should adopt a discretionary and more flexible central bank policy.

The Role and Functioning of Central Banks Today

Theoretically speaking, there exist two types of money in our financial system: base money and commercial bank money. Base money, the core of the money supply, is created by central banks whenever they buy assets or extend credit to commercial banks. It can be destroyed when central banks sell assets or when credit is repaid. According to Belke and Polleit (2009, p. 29), "[c]ommercial banks need base money for at least three reasons: (i) making inter-bank payments; (ii) meeting any *cash drain* as non-banks want to keep a portion of their deposits in cash (notes and coins); and (iii) holding a certain portion of their liabilities in the form of base money with the central bank (*minimum reserves*)." The minimum reserve requirement is one instrument of monetary policy to regulate the overall money supply in the economy. Minimum reserves are held in cash directly at the commercial bank or as deposits with the corresponding central bank. Given a reserve requirement of 1% as currently in the eurozone, for any unit of base money, a commercial bank can create commercial bank money by extending loans of up to 99 units to the private sector (*excess reserves*). Hence, the maximum money creation potential of the banking sector is determined by the central bank through minimum reserve requirements.

The primary tools of central banks to manage the supply of base money are open market operations (Belke and Polleit 2009, p. 33). Their implementations work similarly in the United States and in the eurozone. However, minor differences can be observed. In the United States, the *Domestic Trading Desk* (Desk) arranges open market operations on a daily basis. It has to figure out whether there are imbalances between supply and demand for base money and react accordingly. Imbalances are indicated by differentials between the effective interest rate at which base money is borrowed and lent on the interbank market and the target interest rate set by the *Federal Open Market Committee*. Usually, short-lived imbalances are corrected through *temporary* operations. In special cases, when imbalances turn out to be more persistent than expected, the Desk may perform *outright* operations. These operations involve the buying and selling of government bonds on the *secondary market*, that is, the market in which formerly issued government bonds are traded. This means that, under normal circumstances, the FED does not buy bonds directly from the government. Outright open market operations affect the base money supply permanently, whereas the much more common temporary open market operations will unwind after a specified number of days. The latter are combined with *repurchase agreements*. For example, the Desk may decide to increase the base money supply and buys government securities from commercial banks. In order to keep this increase temporary, it agrees to resell those securities to its counterparties on a future date. *Matched sale-purchase transactions* are the tool with which the base money supply is temporarily decreased. Securities are first sold and will then be bought back in the future. The Desk may also redeem maturing securities, rather than replacing them with new ones, and can thereby reduce the portfolio without entering the market directly (Edwards 1997).

Similar to the Desk, the ECB mostly uses reverse transactions, that is, buying or selling eligible assets under repurchase agreements or lending money against eligible assets provided as collateral (Belke and Polleit 2009, p. 43). These transactions are used for *main refinancing operations* with a maturity of usually one week as well as *longer-term refinancing operations* with a maturity of usually three months. The ECB may use *fine-tuning operations* in reaction to unexpected liquidity fluctuations to steer interest rates in the form of *outright transactions* and *foreign exchange swaps.* The latter are spot and forward transactions in euro against foreign currencies. Furthermore, the ECB manages its structural position vis-à-vis the financial sector by issuing *ECB debt certificates* (ECB 2006). One of the main policy objectives of the ECB is to control short-

term interest rates and reduce their volatilities. The *marginal lending facility* and the *deposit facility* are always available for credit institutions on their own initiative, whenever there is a lack of trading partners on the money market. Interest rates at the lending facility are usually higher than on the money market. The deposit facility usually offers lower interest rates. Those two institutions therefore provide boundaries within which interest rates on the overnight money market fluctuate. There is no limit on the access to these facilities other than collateral requirements at the lending facility.

Central banks generally have to decide on which policy instruments to use. In a simplified version, the problem can be seen as a choice between setting interest rates and letting the money supply be determined endogenously or the other way around. In practice, as Blinder (1998) points out, there are many more choices to be made, including various definitions of the money supply, several different choices for interest rates, bank reserves, and exchange rates. The practical problems involved might be more complicated than the theory suggests.

> The intellectual problem is straightforward in principle. For any choice of instrument, you can write down and solve an appropriately complex dynamic optimization problem, compute the minimized value of the loss function, and then select the *minimum minimorum* to determine the optimal policy instrument. In practice, this is a prodigious technical feat that is rarely carried out. And I am pretty sure that no central bank has ever selected its instrument this way. But, then again, billiards players may practice physics only intuitively. (Blinder 1998, pp. 26–27)

As the above quote suggests, there is usually a clear discrepancy between the theoretically optimal and the practically possible. One of the various controversial subjects, when it comes to monetary policy in practice, is the question whether political actions should be rule based or discretionary. It has been argued that central banks left with discretion tend to err systematically in the direction of too much inflation. In order to correct this bias, one needs more or less strict rules (Kydland and Prescott 1977). Simons (1936) favored a strict commitment of the FED to price stability, that is, zero inflation, rather than pursuing any other possible policy goals. Friedman (1959) and other economists in the monetarist tradition advocated a low but constant growth rate of the money stock. Yet another, even more restrictive, rule was proposed by Wallace (1977). He argued that the FED should consider holding the money stock constant. Such a rule would essentially end monetary policy altogether. This view has not found many adherents. Instead, a modern trend in central banking is inflation rate targeting that only evolved after the end of the Bretton Woods system, in which exchange rates were targeted. Usually, inflation targets are given in a more or less narrow range around 2%. Allowing some inflation to take place within a certain range provides more flexibility to pursue other policy goals. The target can be met through adjusting interest rates appropriately. Whenever inflation rates are below the target range, interest rates can be lowered and vice versa. A declared inflation target also makes central bank policy more transparent. Investors can more easily anticipate possible changes in interest rates. This may lead to an overall stabilization of the economy. However, in the course of the current financial crisis (since 2008), inflation targeting has been abandoned in order to intervene more actively into the economy by means of more expansionary monetary policy. Interest rates are lower than ever before, providing liquidity for credit institutions on the financial markets at the risk of higher inflation rates. The reactions on the current crisis have been criticized on very different grounds. For some economists, they are still too conservative. For others, they constitute only a treatment of the symptoms, rather than the causes. In their view, they will prolong the problem instead of solving it. Central banking practices in general have been criticized, both on purely economical and ethical grounds.

Critique of Central Banking

A first very general point of criticism lies in the state-granted monopoly status of central banks. What justifies a legal monopoly over the supply

of money? There has been a lively debate over the question whether money is a natural monopoly. For this to be the case, according to the traditional definition, the production of money should exhibit economies of scale, which means that the average costs of production for one firm producing the whole output are always lower than if two or more firms with access to the same technology divide the output (Glasner 1998, p. 23). However, the fact that central banks under today's fiat money regime produce money at almost zero production costs does not imply economies of scale. Anyone could produce an unbacked or digital money at almost zero production costs.

The demand side characteristics of money also fall short of justifying its legal monopoly. Even if it is theoretically beneficial and cost reducing to use only one universally accepted medium of exchange within a given area and even if in fact only one universally accepted medium of exchange would emerge among freely cooperating individuals, for example, gold, there is no justification for legally restricting the production of that medium to one authorized institution. Evidently, for the functioning of a fiat money regime, it is necessary to restrict production to one institution, since otherwise competition would lead to an excessive expansion of the money supply and a rapid devaluation until its value reaches production costs – essentially zero (Hoppe 2006). But as mentioned above, the fiat money regimes that govern today's global economy are themselves the result of state interventions into traditional commodity standards. Therefore, the mere existence of a fiat money regime cannot justify the legal monopoly per se. Furthermore, to consider money to be a public good is false, since it lacks the criteria of non-rivalness and non-excludability (Vaubel 1984). If, however, the money production is legally monopolized by establishing central banks, then the general economic analysis of monopolies should be applicable to it. In general, monopolies are considered to be economically inefficient and costly. There is an omnipresent danger that the monopoly status is irresponsibly exploited at the expense of the public.

Central banks in a fiat money regime are able to create as much money as they please and can serve as a l*ender of last resort* for banks that lack liquidity. Within such an environment, an incentive for commercial banks is set to operate under a lower equity ratio, to hold less money, and to engage in riskier projects, as they can always borrow money at relatively low interest rates from the central bank. Hence, a *moral hazard* problem arises. As a result, the financial system as a whole becomes more fragile and susceptible to crises. The same analysis holds for governments. The European debt and financial crisis is a dramatic case in point (Bagus 2012).

Opposed to the Keynesian theory that ultimately monetary spending determines economic progress, the *Austrian Theory of the Business Cycle* advises against artificial credit expansion through lowering interest rates, which is easier to accomplish than ever before under a fiat money regime controlled by central banks. This theory, introduced by Ludwig von Mises (Mises 1912) and further developed by Friedrich August von Hayek (Hayek 1935), sees the root cause of economic depressions in an imbalance between investments and real savings brought about through this very process of credit expansion. In the Austrian view, the interest rate is not an arbitrary number that should be interfered with. Instead, it is the price that tends to accommodate the roundaboutness of production processes or investment projects to the available subsistence fund in the economy. Interest rates tend to fall when consumers save more and thereby increase the subsistence fund. In these situations, more roundabout investment projects can be sustained. If, however, interest rates are lowered artificially, the subsequent excess investments are not covered by real savings. At least some of the *malinvestments* have to be liquidated when the imbalance becomes evident. This situation constitutes the economic bust.

Another point of criticism lies in the inflationary tendencies of the modern monetary system under central bank control. Hyperinflation rates of more than 100%, as in the Weimar Republic or more recently in Zimbabwe, have obviously devastating effects. But also, moderate inflation rates are not neutral. They can be interpreted as a tax

that enables governments to pursue policy goals that lack democratic legitimization (Hülsmann 2008, p. 191). State control over money was the result of the "characteristic quest by the state for sources of revenue" (Glasner 1998, p. 24). Fiat inflation therefore leads to an excessive growth of the state for which the citizens do not pay directly through taxes but rather indirectly through a devaluation of the currency they use. The redistributive effects of inflation are known as Cantillon effects (Cantillon 2010). Since inflation does not take place uniformly, but rather gradually ripples through the economy, the first receivers of the newly created money benefit on the expense of all others, since they can buy goods on the market for still relatively low prices. As they spend the money, prices tend to rise. Late receivers and people on fixed incomes face price increases before or entirely without increases in their incomes. They suffer a loss in real terms. Usually, the first receivers and beneficiaries of newly created money are commercial banks, other financial institutions, governments, and closely related industries. It is argued that the general public carries the burden. Although some groups doubtlessly benefit from the inflationary tendencies of the fiat money system, some economists argue that it cannot benefit society as a whole. The mere possibility to position oneself on the winner side leads to some kind of "collective corruption" and the maintenance of the system (Polleit 2011).

References

Bagus P (2012) The tragedy of the Euro, 2nd edn. Ludwig von Mises Institute. Available Online: http://mises.org/document/6045/The-Tragedy-of-the-Euro

Belke A, Polleit T (2009) Monetary economics in globalised financial markets. Springer, Berlin/Heidelberg

Blinder AS (1998) Central banking in theory and practice (the Lionel Robbins lectures). Massachusetts Institute of Technology, Cambridge

Bordo MD (2007) A brief history of central banks, Federal Reserve Bank of Cleveland – economic commentary. Available Online: http://www.clevelandfed.org/research/commentary/2007/12.cfm

Bordo MD, Kydland FE (1995) The gold standard as a rule: an essay in exploration. Explor Econ Hist 32:423–464

Cantillon R (2010) An essay in economic theory. Ludwig von Mises Institute. Available Online: http://mises.org/document/5773/An-Essay-on-Economic-Theory

de Roover R (1976) Chapter 5: New interpretations in banking history. In: Business, banking, and economic thought in the Middle Ages and early modern Europe. University of Chicago Press

ECB (2006) The implementation of monetary policy in the Euro area, general document of eurosystem monetary policy instruments and procedures. Available Online: http://www.ecb.europa.eu/pub/pdf/other/gendoc2006en.pdf

Edwards CL (1997) Open market operations in the 1990s. Federal Reserve Bulletin, pp 859–874. Available Online: http://www.federalreserve.gov/pubs/bulletin/1997/199711lead.pdf

Flandreau M, Maurel M (2001) Monetary union, trade integration and business fluctuations in 19th century Europe: just do it. Centre for Economic Policy Research working paper no. 3087

Friedman M (1959) A program for monetary stability. Fordham University Press, New York

Glasner D (1998) An evolutionary theory of the state monopoly over money, in money and the nation state – the financial revolution, government and the world monetary system. The Independent Institute, Oakland

Hazlitt H (1959) The failure of the "New Economics" – an analysis of the Keynesian Fallacies. D. van Nostrand. Available Online: http://mises.org/document/3655/Failure-of-the-New-Economics

Hoppe H-H (2006) How is fiat money possible? – or, the devolution of money and credit, published in the economics and ethics of private property – studies in political economy and philosophy, 2nd edn. Ludwig von Mises Institute. Available Online: http://mises.org/document/860/Economics-and-Ethics-of-Private-Property-Studies-in-Political-Economy-and-Philosophy-The

Hülsmann JG (2008) The ethics of money production. Ludwig von Mises Institute. Available Online: http://mises.org/document/3747/The-Ethics-of-Money-Production

Keynes JM (1936) The general theory of employment, interest and money, Macmillan Cambridge University Press. A revised version is Available Online: http://www.marxists.org/reference/subject/economics/keynes/general-theory/

Krugman P (2006) Introduction to the general theory of employment, interest, and money, by John Maynard Keynes. Available Online: http://www.pkarchive.org/economy/GeneralTheoryKeynesIntro.html

Kydland FE, Prescott EC (1977) Rules rather than discretion: the inconsistency of optimal plans. J Polit Econ 85(3):473–492

Lane FC (1937) Venetian bankers, 1496–1533: a study in the early stages of deposit banking. J Polit Econ 45(2):187–206

López-Córdova JE, Meissner CM (2003) Exchange-rate regimes and international trade: evidence from the classical gold standard era. Am Econ Rev 93(1): 344–353
von Mises L (1912) Theorie des Geldes und der Umlaufsmittel. Verlag von Duncker und Humblot, München und Leipzig. Available Online: http://mises.org/document/3298/Theorie-des-geldes-und-der-Umlaufsmittel; for the English edition see: http://mises.org/document/194/The-Theory-of-Money-and-Credit
Polleit T (2011) Fiat money and collective corruption the quarterly. J Aust Econ 14(4): 297–415. Available Online: https://mises.org/journals/qjae/pdf/qjae14_4_1.pdf
Prestwich M (1979) Italian merchants in late thirteenth and fourteenth century england, in the dawn of modern banking. Yale University Press, New Haven
Romer CD (2003) Great depression, forthcoming in the encyclopædia britannica Available Online: http://eml.berkeley.edu/~cromer/great_depression.pdf
Rothbard MN (1998) The gold exchange standard in the interwar years, in money and the Nation State – The Financial Revolution, Government and the World Monetary System, The Independent Institute, Oakland. Also published in A history of money and banking in the United States – the Colonial Era to World War II and Available Online: http://mises.org/document/1022/History-of-Money-and-Banking-in-the-United-States-The-Colonial-Era-to-World-War-II
Simons HC (1936) Rules versus authorities in monetary policy. J Polit Econ 44(1):1–30
von Hayek FA (1935) Prices and production, 2nd edn. Augustus M. Kelley, New York Available Online: http://mises.org/document/681/Prices-and-Production
Usher AP (1943) The early history of deposit banking in mediterranean Europe. Harvard University Press, Cambridge
Vaubel R (1984) The Government's money monopoly: externalities or natural monopoly. Kyklos 37(1): 27–58
Wallace N (1977) Why the fed should consider holding M0 constant. Q Rev Federal Reserve Bank of Minneapolis, Minneapolis, MN 1(1):2–10

Certification Labeling

▶ Labeling

Change of Circumstances

▶ Impracticability

Child Maltreatment, The Economic Determinants of

Jason M. Lindo[1,2,3] and Jessamyn Schaller[4]
[1]Texas A&M University, College Station, TX, USA
[2]NBER, Cambridge, MA, USA
[3]IZA, Bonn, Germany
[4]The University of Arizona, Tucson, AZ, USA

Abstract

This entry examines the economic determinants of child maltreatment. We first discuss potential mechanisms through which economic factors, including income, employment, aggregate economic conditions, and welfare receipt, might have causal effects on the rates of child abuse and neglect. We then outline the main challenges faced by researchers attempting to identify these causal effects, emphasizing the importance of data limitations and potential confounding factors at both the individual and aggregate levels. We describe two approaches used in the existing literature to address these challenges – the use of experimental variation to identify the effects of changes in family income on individual likelihood of maltreatment and the use of area studies to identify the effects of changes in local economic conditions on aggregate rates of maltreatment.

Definition

The economic determinants of child maltreatment refer to the broad set of economic factors that have causal effects on the rates of child abuse and neglect, either directly or indirectly, potentially including income, employment, aggregate economic conditions, and welfare receipt.

Introduction

Child maltreatment, including physical abuse, sexual abuse, emotional abuse, and neglect, is a

prevalent and serious problem. In the United States alone, more than six million children are involved in reports to Child Protective Services (CPS) annually, while countless more are subject to unreported maltreatment (Petersen et al. 2014). Child maltreatment has severe and lasting consequences for victims, injuring physical and mental health and affecting interpersonal relationships, educational achievement, labor force outcomes, and criminal behavior (see, e.g., Gilbert et al. 2009; Berger and Waldfogel 2011). Child maltreatment is costly to society as well, generating productivity losses, increased burdens on criminal justice systems and special education programs, and substantial costs for child welfare services and health care (Fang et al. 2012; Gelles and Perlman 2012).

Given the pervasive and damaging nature of the problem, it is not surprising that a substantial literature spanning many disciplines and several decades is devoted to identifying the causes of child maltreatment. (For a summary of this literature, see Petersen et al. (2014).) Within this literature, a variety of economic factors, including family income, parental employment, macroeconomic conditions, and welfare receipt, have been identified as predictors of child abuse and neglect (Pelton 1994; Stith et al. 2009; Berger and Waldfogel 2011). Yet, due to data limitations and identification challenges, researchers have only recently begun to make progress isolating the causal effects of these variables on maltreatment.

This entry is devoted to the economic determinants of child maltreatment. We begin with etiological theories of child maltreatment from the fields of psychology and economics, outlining the potential mechanisms by which different economic factors might be correlated with child abuse and neglect at the individual and aggregate levels. Next, we describe different types of data used in the study of child maltreatment and discuss their limitations. We then discuss the additional challenges that maltreatment researchers face in estimating the causal effects of economic conditions, the empirical approaches that researchers have taken to try to overcome these challenges, and the lessons learned from these studies before concluding.

Theory and Mechanisms

The most commonly cited etiological models of child maltreatment are the developmental-ecological and ecological-transactional models originating in psychology (Garbarino 1977; Belsky 1980; Cicchetti and Lynch 1993). These models posit that maltreatment results from complex interactions between individual, familial, environmental, and societal risk factors. Among the risk factors for maltreatment in these models, economic variables, such as family income and parental employment status, have garnered particular attention in the literature, both because they are robust, easily measured predictors of maltreatment and because they can be manipulated through policy intervention. However, as ecological models posit that maltreatment results from *interactions* between economic variables and characteristics of individuals, families, and communities, these models do not generate clear predictions about how economic factors should be correlated with maltreatment. (For example, the effect of a stressful life event such as a reduction in family income on the likelihood of maltreatment may be exacerbated by individual characteristics such as depression while also being mitigated by social support and other buffering factors (National Research Council 1993).)

Economists have approached theoretical modeling of child maltreatment from a different perspective, seeking to understand child maltreatment within a framework of budget constraints and utility functions. Several empirical investigations of child maltreatment, including those of Paxson and Waldfogel (2002), Seiglie (2004), Berger (2004, 2005), and Lindo et al. (2013), have been motivated by theoretical models of investments in child quality, sometimes in combination with altruistic, cooperative bargaining, and noncooperative bargaining models used in economic studies of marriage and divorce, family labor supply, and domestic partner violence. There is also overlap between theoretical models of child maltreatment and economic models of criminal behavior. (Berger (2004, 2005) provides a nice summary of several theoretical economic models relevant to the analysis of child abuse and

neglect. To our knowledge, the only study with formal model of child maltreatment is Seiglie (2004), which builds on economic models of investment in child quality.)

In developing a theoretical framework for understanding the oft-observed link between poverty and maltreatment, it is important to distinguish between reasons child maltreatment might be associated with poverty and causal pathways through which economic variables might affect the incidence of abuse and neglect. For example, parental education, community norms with regard to parenting behaviors, parental history of abuse, and innate personality characteristics of parents have all been cited as important factors that could explain some (or potentially all) of the association between poverty and child maltreatment. In thinking about the *causal pathways* through which economic factors may affect child maltreatment, it may be useful to imagine a hypothetical experiment in which a household is randomly selected to receive an intervention such as a cash transfer, an unanticipated job displacement, or a change in aggregate economic conditions and to consider the effects of this treatment on the likelihood that the children in that household will experience abuse or neglect. With these types of experiments in mind, researchers have identified a number of potential pathways through which these economic "treatments" might influence the likelihood of child abuse and neglect. (In this section we focus on the relationship between economic factors and the likelihood of committing maltreatment rather than the likelihood of being reported, investigated, or punished for abuse. We discuss issues related to reporting and data quality in the next section.)

First, income may have direct effects on the likelihood of maltreatment if parents are constrained in their ability to provide sufficient care for their children (Berger and Waldfogel 2011). This mechanism is particularly relevant to the study of child neglect, which is commonly defined as the failure of a caregiver to provide for a child's basic physical, medical, educational, or emotional needs, and thus is often considered to be "underinvestment" in children within the context of economic models (see, e.g., Seiglie 2004). (Weinberg (2001) notes that family income may be directly associated with abuse as well, as it relates to the availability of resources that can be used to elicit desired behavior from children.)

Changes in the amount and sources of family income may also affect child maltreatment by altering the distribution of bargaining power within households and changing the expected cost of abuse. Building on bargaining models used in economic studies of domestic violence, Berger (2005) posits that, in two-parent households, shifts in the distribution of family income away from the perpetrator of abuse and toward a non-abusing partner can result in a shift in the balance of power within the relationship, which can in turn affect the incidence of maltreatment. Additionally, as in economic models of criminal behavior, income shocks can affect the cost that the perpetrator of maltreatment expects to incur if he/she is caught. Specifically, the perpetrator's access to income is jeopardized if maltreatment leads to dissolution of a relationship and loss of access to a partner's income. The removal of a child can also lead to the loss of child-conditioned transfers such as welfare payments and child support.

Economic shocks may also affect rates of child abuse and neglect through their impacts on mental health. At the aggregate level, research has shown that economic downturns are associated with deterioration of population mental health, as measured by the incidence of mental disorders, admissions to mental health facilities, and suicide (Zivin et al. 2011). Job displacement has also been linked to a number of mental-health-related outcomes, including psychological distress (Mendolia 2014), depression (Brand et al. 2008; Schaller and Stevens 2014), psychiatric hospitalization (Eliason and Storrie 2010), and suicide (Eliason and Storrie 2009; Browning and Heinesen 2012). Meanwhile at the individual level, a large literature documents a correlation between poverty and mental health in the cross-section. However, empirical evidence on the causal effects of individual and family income on mental health is sparse and inconclusive. (Several papers have examined mental health outcomes of lottery winners, with mixed results (e.g., Kuhn et al. 2011; Apouey and Clark 2014).)

Substance abuse and partnership dissolution may also mediate the relationship between economic shocks and child maltreatment. Alcohol and drug use and single parenthood are both correlated with socioeconomic status and are also well-known risk factors for child abuse and neglect. However, the causal links between economic shocks and these variables are not well understood. (For example, Deb et al. (2011) identify heterogeneity in the response of drinking behavior to job displacement and the empirical evidence on the effects of aggregate economic downturns on alcohol consumption is mixed (Ruhm and Black 2002; Dávalos et al. 2012). Meanwhile, while layoffs lead to increased divorce rates in survey data (Charles and Stephens Jr 2004; Doiron and Mendolia 2012), aggregate divorce rates are found to decrease in recessions (Schaller 2013).)

Finally, parental time use is a rarely mentioned mechanism by which economic shocks can affect maltreatment. In particular, involuntary changes in employment and work hours have the potential to affect the incidence of maltreatment through their effects on the amount of time children spend with parents, other family members, childcare providers, and others (Lindo et al. 2013). This mechanism may work in different directions depending which parent experiences the employment shock and on the type of maltreatment considered. (A shock that shifts the distribution of childcare from the mother to the father may increase the incidence of abuse since males tend to have more violent tendencies than females. As another example, additional time at home with a parent may reduce the likelihood of child neglect but increase the likelihood of physical, sexual, and emotional abuse.)

Identifying Causal Effects

Identifying the causal effects of economic factors on child maltreatment requires (i) child maltreatment data linked to measures of economic conditions and (ii) empirical strategies that can isolate the effects of economic factors despite the fact that these factors tend to be correlated with other determinants of maltreatment. Both of these issues present challenges for researchers that are difficult – though not impossible – to overcome.

Data

Data Based on Maltreatment Reports

Child abuse reports have historically been the primary source of data for researchers interested in studying child maltreatment on a large scale. While these data are attractive because they often span large areas and many time periods, a natural concern is that maltreatment report data may not accurately reflect the true incidence of maltreatment. While there is no doubt that false reports are sometimes made, the consensus view is that statistics tend to understate the true prevalence of child abuse because underreporting is such a serious issue (Waldfogel 2000; Sedlak et al. 2010). In fact, the Fourth National Incidence Study of Child Abuse and Neglect (NIS-4), which identifies maltreated children outside of the United States Child Protective Services (CPS) system, found that CPS investigated the maltreatment of only 32% of children identified in the study as having experienced observable harm from maltreatment. Applying CPS screening criteria to the maltreatment cases that were not investigated by CPS, the researchers concluded that underreporting was the primary reason for this low rate of investigation: three quarters of the cases would have been investigated if they had been reported to CPS (Sedlak et al. 2010).

Nonetheless, reports are likely to be strongly related to the true incidence of maltreatment and thus may serve as a useful proxy. The key consideration with the use of any proxy variable is the degree to which the measurement error is the same across comparison groups. If a comparison is made across groups that have the same degree of measurement error (or across time periods that have the same degree of measurement error), then the percent difference in the proxy will be identical to the percent difference in the variable of interest. For example, if State A has 1,200 maltreatment reports and State B has 800 maltreatment reports and the true incidence of maltreatment is understated in both states by 20%, then the

percent difference in reports ((1,200–800)/800 × 100% = 50%) will be equal to the percent difference in the true incidence of maltreatment (1,200 × 1.2 − 800 × 1.2)/800 × 1.2 × 100% = 50 %).

Given that estimating the causal effects of economic factors on child maltreatment will inevitably entail comparisons across groups and/or time periods, this discussion naturally raises the question: is it generally safe to assume that the measurement error in abuse reports is the same across groups and across time? Unfortunately for researchers, while this assumption may hold in certain circumstances, it is unlikely to hold in most instances. When making comparisons across states, we must address the fact that states differ in how they define abuse, who is required to report abuse, and in how they record and respond to reports of abuse. When making comparisons across time, we must acknowledge that children's exposure to potential reporters and individual propensities to report maltreatment may be changing over time and that the rate of reporting may in fact even be correlated with economic factors. Moreover, states have periodically changed their official definitions of abuse, reporting expectations, and standards for screening allegations. As such, comparisons of abuse reports across states and time have the potential to reflect differences in measurement error in addition to differences in the incidence of maltreatment. Comparisons across groups defined in other ways will be susceptible to similar issues. For example, the maltreatment of infants and toddlers may be less likely to be detected than the maltreatment of school-aged children who spend more time in the presence of mandatory reporters.

It is also important to note that focusing on substantiated reports does not necessarily improve our ability to make valid comparisons – and could actually make things worse – even in a scenario in which agencies are perfectly able to discern true and false reports. Comparisons of substantiated reports (in percent terms) will do better than comparisons of all reports if and only if the *difference* in the measurement error in substantiated reports across groups is less than the *difference* in the measurement error in overall reports across groups, which may not be the case. (Here the measurement error we refer to is the degree to which the variable differs from what we would like to measure: true incidents. As an example in which we would do worse by focusing on substantiated reports, suppose State C has 2,500 true incidents, 40% of which are reported, and 5 false reports per 100 true incidents, while State D has 2,000 incidents, 35% of which are reported, and 10 false reports per 100 true incidents. Then, assuming true reports are substantiated and false reports are not substantiated, the percent difference in reports would correctly identify the true percent difference in incidents, whereas the percent difference in substantiated reports would not, as the true percent difference = (2,500–2,000)/2,000 × 100% = 25%, the percent difference in reports = [2,500 × (40% + 5%) − 2,000 × (35% + 10%)]/2,000 × (35% + 10%) × 100% = 25%, and the percent difference in substantiated reports = (2,500 × 40% − 2,000 × 35%)/2,000 × 35% × 100% = 43%.)

The major take-away from this discussion is that we must take into consideration the process by which maltreatment that occurs becomes observable to the researcher. In particular, when a researcher estimates the causal effect of an economic factor on the observed incidence of maltreatment, we must consider the degree to which the effects are driven by actual changes in maltreatment and/or by changes in the rate at which occurrences of maltreatment are detected and reported.

Alternative Sources of Data

Survey data, hospital data, and internet search data have also been used to gain insights into the prevalence of maltreatment and the way it varies with economic factors. Cross-sectional surveys include retrospective questionnaires that solicit information on occurrences of maltreatment over one's childhood or within a specific time window, while panel surveys solicit information on a year-to-year basis. Hospital data can be used to measure maltreatment using diagnosis codes that explicitly indicate maltreatment or by considering outcomes that are expected to be highly correlated with maltreatment (e.g., accidents, shaken-baby syndrome, etc.), as in Wood et al. (2012). And

internet search data can be used to measure the frequency with individuals are searching for phrases that are expected to be highly correlated with maltreatment (e.g., child protective services, dad hit me, etc.), as in Stephens-Davidowitz (2013).

While all of these sources of data have the potential to shed new light on maltreatment in ways that administrative reports data cannot, they are also susceptible selection bias. Just as economic factors may affect both the incidence of maltreatment and the likelihood that maltreatment cases are reported to officials, economic factors may affect the likelihood that an individual reports being abused in a questionnaire, the likelihood that a doctor's diagnosis involves maltreatment, the likelihood that a maltreated child is taken to the hospital, or the likelihood that individuals suspecting or experiencing maltreatment search the internet for information. As such, they do not lessen the importance of considering the process by which maltreatment that occurs becomes observable to the researcher.

Links to Measures of Economic Conditions

Because of the sensitive nature of the subject, most maltreatment data are only available as aggregates (e.g., counts for states and years). Where micro data is available, it often does not include information on families' economic circumstances. As such, it is often only possible to consider links between maltreatment and the economic conditions of an area, which introduces the possibility that estimated relationships may be subject to the ecological fallacy. That is, a relationship between economic conditions and maltreatment that is observed in the aggregate may not reflect the relationship that exists for individuals. For example, it is possible for unemployment at the local level to increase child maltreatment while an individual being unemployed may have the opposite effect. Nonetheless, while it is important to acknowledge the limitations of what can be learned from estimates based on aggregate data, it is also important to note that there is value to understanding the links between economic conditions and child maltreatment in the aggregate.

With that said, some data on child maltreatment *do* provide information on the economic conditions of the household that the child lives in. It is from these data that we know that maltreated children tend to come from households that are economically disadvantaged relative to the average household. While these sorts of data are useful for providing descriptive statistics for children who are (observed) maltreated, data that has been selected on the outcome of interest cannot be used estimate causal links in any straightforward manner. Using microlevel data to estimate the degree to which various factors affect the probability of maltreatment requires data on individuals who are *not* maltreated in addition to those who are maltreated. Toward this end, researchers have used survey data including the National Family Violence Survey, the Fragile Families and Child Wellbeing Study, the National Longitudinal Survey of Youth, and by linking data sets with information on economic conditions to child abuse report data.

Empirical Strategies

As discussed in the "Theory and Mechanisms" section above, child maltreatment can be thought of as resulting from complex interactions between individual, familial, environmental, and societal risk factors. Given the large number of factors that may contribute to maltreatment and the interrelatedness of these factors, researchers face a major challenge in trying to identify the causal effects of economic conditions on maltreatment. In this section we highlight two approaches to overcoming this challenge, one that is best suited for estimating the effects of household economic factors and one that is best suited for estimating the effects of broader economic conditions.

Estimating the Effects of Household Economic Factors

Acknowledging that household economic conditions are generally *not* random, quantifying their causal effects requires researchers to consider circumstances in which they can measure the effects of random shocks to these conditions. Because it is difficult to identify these circumstances and to collect the maltreatment data necessary to

examine these circumstances, only a handful of such studies exist.

Fein and Lee (2003) take this approach in an experimental evaluation of a welfare reform program in Delaware. In particular, they compare outcomes for households subject to welfare reform to outcomes for those who were not subject to welfare reform, which was determined by random assignment. They find that the reform increased the incidence of reports of neglect but had no significant effect on reports of abuse or foster care placement. While this study represents some of the most convincing evidence to date that household economic factors have a causal effect on child maltreatment, it also underscores the difficulty of teasing out the causal effects of different interrelated economic factors. In particular, Delaware's welfare reform involved changes to benefit levels and work incentives in addition to other factors, any of which may have contributed to the increase in reports of neglect.

Cancian et al. (2013) also exploit evidence based on an experiment among welfare recipients to learn about the causal effect of household income on child maltreatment. In particular, they evaluate the effect of Wisconsin's reform that allowed a full pass-through of child support to welfare recipients (as opposed to the prior policy in which the government retained a fraction of child support payments to offset costs). Because the experimental intervention only changed child support pass-through – and no other aspect of child support or welfare receipt – the design allows for a straightforward interpretation of the results: that increasing income through this mechanism reduces maltreatment reports. The authors are careful to note, however, that increasing income through other mechanisms may have different effects on maltreatment. For example, an increase in income that is generated by an increase in maternal labor supply could very well increase the incidence of maltreatment.

Berger et al. (2014) take a different approach to identifying the causal effect of household economic conditions, exploiting naturally occurring variation in income (as opposed to experimentally manipulated variation) that they argue can be thought of as random. In particular, their strategy utilizes variation in the generosity of the Earned Income Tax Credit (EITC) across states and over time. While this approach allows for a study that is broader in scope than the aforementioned experiments, a disadvantage of this approach is that changes in EITC rules can affect levels of income, work activity, and the broader social economic climate, which again highlights the challenge in the identification and interpretation of causal effects.

Estimating the Effects of Broader Economic Conditions

Another strand of the literature on the causal effects of economic conditions on child maltreatment abstracts from the household to consider the effects of changes in local economic conditions on rates of maltreatment in the aggregate. Acknowledging that local economic conditions tend to be correlated with many socioeconomic factors that predict maltreatment, several studies have taken an "area approach" that considers how rates of maltreatment in an area change *over and above changes occurring across all areas* when its economic conditions change *over and above changes occurring across all areas*. As such, estimates based on this approach are identified using variation across areas in the timing and severity of changing economic conditions. This approach is operationalized via regression models that include time-fixed effects to capture changes occurring across all areas at the same time, area-fixed effects to capture time-invariant area characteristics, and (sometimes) area-specific trends. The validity of this approach rests on the assumption that unobservable variables related to the outcome variable do not deviate from an area's trend when its economic conditions deviate from trend.

Studies taking this approach vary considerably in their measures of maltreatment, their measures of economic conditions, and the way they define areas. Paxson and Waldfogel (1999, 2002, 2003), Seiglie (2004), and Bitler and Zavodny (2002, 2004) use state-level panel data to estimate the effects of a variety of economic indicators on maltreatment reports, finding mixed results. Lindo et al. (2013) and Frioux et al. (2014) use county-level data from California and

Pennsylvania, respectively, also finding mixed results. Wood et al. (2012) focus on hospital admissions for abuse-related injuries using panel data from 38 hospitals from 2000 to 2009 along with a variety of economic indicators and find evidence that local economic downturns significantly increase the incidence of severe physical abuse; however, they do not account for the likely autocorrelation in the error terms within hospitals over time, which would serve to widen their confidence intervals.

Conclusion

Child maltreatment is an important topic that has received relatively little attention in the field of economics, despite generating large financial costs for society and significant consequences for the health, human capital accumulation, and eventual labor market outcomes of its victims. The scarcity of economic research on the topic is especially unfortunate given that a literature spanning many disciplines and several decades has found economic factors, including local economic conditions, family income, neighborhood poverty, employment status, and receipt of public assistance, to be robust predictors of child abuse. We suspect that this scarcity is driven by economists' strong emphasis on the identification of causal effects, which is particularly challenging for research on the economic determinants of child maltreatment. In some sense, identifying causal effects in this area requires a perfect storm in which there is random variation in economic conditions, the researcher has access to maltreatment data that allows for comparisons utilizing this random variation, and the researcher can be confident that the way in which maltreatment becomes observed in these data does not vary across the groups of individuals and/or time periods he/she intends to compare. Moreover, even when this perfect storm occurs such that a causal estimate can be obtained, the interrelatedness of economic factors can make it difficult to interpret such estimates. For example, the causal effect of a parent's job displacement could reflect the effects of income or time use (or other factors). Despite these challenges, recent progress has been made in identifying the causal effects of economic factors on child maltreatment through the use of experimental (natural and true) variation and area studies. These studies indicate that changes in economic conditions can have meaningful impacts on maltreatment. However, there is still much work to be done in identifying exactly which economic factors matter and in characterizing the nature of these relationships.

C

References

Apouey B, Clark AE (2014) Winning big but feeling no better? The effect of lottery prizes on physical and mental health. Health Econ, http://onlinelibrary.wiley.com/doi/10.1002/hec.3035/full

Belsky J (1980) Child maltreatment: an ecological integration. Am Psychol 35(4):320–335

Berger LM (2004) Income, family structure, and child maltreatment risk. Child Youth Serv Rev 26(8): 725–748

Berger LM (2005) Income, family characteristics, and physical violence toward children. Child Abuse Negl 29(2):107–133

Berger LM, Waldfogel J (2011) Economic determinants and consequences of child maltreatment. OECD social, Employment and migration working papers, No. 111, http://storage.globalcitizen.net/data/topic/knowledge/uploads/20120229105523533.pdf

Berger Lawrence M, Sarah A Font, Kristen S Slack, Jane Waldfogel (2014) Income and child maltreatment: evidence from the earned income tax credit Mimeo

Bitler M, Zavodny M (2002) Child abuse and abortion availability. Am Econ Rev 92(2):363–367

Bitler M, Zavodny M (2004) Child maltreatment, abortion availability, and economic conditions. Rev Econ Househ 2(2):119–141

Brand JE, Levy BR, Gallo WT (2008) Effects of layoffs and plant closings on subsequent depression among older workers. Res Aging 30(6):701–721

Browning M, Heinesen E (2012) Effect of job loss due to plant closure on mortality and hospitalization. J Health Econ 31(4):599–616

Cancian M, Yang M-Y, Slack KS (2013) The effect of additional child support income on the risk of child maltreatment. Soc Serv Rev 87(3):417–437

Charles KK, Stephens M Jr (2004) Job displacement, disability, and divorce. J Labor Econ 22(2):489–522

Cicchetti D, Lynch M (1993) Toward an ecological/transactional model of community violence and child maltreatment: consequences for children's development. Psychiatry 56(1):96–118

Dávalos ME, Fang H, French MT (2012) Easing the pain of an economic downturn: macroeconomic conditions and

excessive alcohol consumption. Health Econ 21(11): 1318–1335
Deb P, William T, Gallo PA, Fletcher JM, Sindelar JL (2011) The effect of job loss on overweight and drinking. J Health Econ 30(2):317–327
Doiron D, Mendolia S (2012) The impact of job loss on family dissolution. J Popul Econ 25(1):367–398
Eliason M, Storrie D (2009) Does job loss shorten life? J Hum Resour 44(2):277–302
Eliason M, Storrie D (2010) Inpatient psychiatric hospitalization following involuntary job loss. Int J Ment Health 39(2):32–55
Fang X, Brown DS, Florence CS, Mercy JA (2012) The economic burden of child maltreatment in the United States and implications for prevention. Child Abuse Negl 36(2):156–165
Fein DJ, Lee WS (2003) The impacts of welfare reform on child maltreatment in delaware. Child Youth Ser Rev 25(1):83–111
Frioux S, Wood JN, Oludolapo F, Luan X, Localio R, Rubin DM (2014) Longitudinal association of county-level economic indicators and child maltreatment incidents. Matern Child Health J, http://link.springer.com/article/10.1007/s10995-014-1469-0
Garbarino J (1977) The human ecology of child maltreatment: a conceptual model for research. J Marriage Fam 39(4):721–735
Gelles RJ, Staci P (2012) Estimated annual cost of child abuse and neglect. Prevent Child Abuse America, Chicago
Gilbert R, Widom CS, Browne K, Fergusson D, Webb E, Janson S (2009) Burden and consequences of child maltreatment in high-income countries. Lancet 373(9657):68–81
Kuhn P, Kooreman P, Soetevent A, Kapteyn A (2011) The effects of lottery prizes on winners and their neighbors: evidence from the Dutch postcode lottery. Am Econ Rev 101(5):2226–2247
Lindo JM, Schaller J, Hansen B (2013) Economic conditions and child abuse. National bureau of economic research working paper w18994, Cambridge, MA
Mendolia S (2014) The impact of husbands job loss on partners mental health. Rev Econ Househ 12(2): 277–294
National Research Council (1993) Understanding child abuse and neglect. The National Academies Press, Washington, DC
Paxson C, Waldfogel J (1999) Parental resources and child abuse and neglect. Am Econ Rev 89(2):239–244
Paxson C, Waldfogel J (2002) Work, welfare, and child maltreatment. J Labor Econ 20(3):435–474
Paxson C, Waldfogel J (2003) Welfare reforms, family resources, and child maltreatment. J Policy Anal Manage 22(1):85–113
Pelton LH (1994) The role of material factors in child abuse and neglect. In: Melton GB, Barry FD (eds) Protecting children from abuse and neglect: foundations for a new national strategy. Guilford Press, New York, pp 131–181
Petersen A, Joshua J, Monica F (eds) (2014) New directions in child abuse and neglect research. The National Academies Press, Washington, DC
Ruhm CJ, Black WE (2002) Does drinking really decrease in bad times? J Health Econ 21(4):659–678
Schaller J (2013) For richer, if not for poorer? Marriage and divorce over the business cycle. J Popul Econ 26(3):1007–1033
Schaller J, Stevens AH (2014) Short-run effects of job loss on health conditions, health insurance, and health care utilization. National bureau of economic research working paper w19884, Cambridge, MA
Sedlak AJ, Mettenburg J, Basena M, Petta I, McPherson K, Greene A, Li S (2010) Fourth national incidence study of child abuse and neglect (NIS-4): report to congress. US Department of Health and Human Services, Administration for Children and Families, Washington D.C.
Seiglie C (2004) Understanding child outcomes: an application to child abuse and neglect. Rev Econ Househ 2(2):143–160
Stephens-Davidowitz S (2013) Unreported victims of an economic downturn. Mimeo, http://static.squarespace.com/static/51d894bee4b01caf88ccb4f3/t/51e22f38e4b0502fe211fab7/1373777720363/childabusepaper13.pdf
Stith SM, Ting Liu L, Davies C, Boykin EL, Alder MC, Harris JM, Som A, McPherson M, Dees JEMEG (2009) Risk factors in child maltreatment: a meta-analytic review of the literature. Aggress Violent Beh 14(1):13–29
Waldfogel J (2000) Child welfare research: how adequate are the data? Child Youth Serv Rev 22(9): 705–741
Weinberg BA (2001) An incentive model of the effect of parental income on children. J Polit Econ 109(2): 266–280
Wood JN, Sheyla P, Medina CF, Luan X, Localio R, Fieldston ES, Rubin DM (2012) Local macroeconomic trends and hospital admissions for child abuse, 2000–2009. Pediatrics 130(2):e358–e364
Zivin K, Paczkowski M, Galea S (2011) Economic downturns and population mental health: research findings, gaps, challenges and priorities. Psychol Med 41(07): 1343–1348

Child Molestation

► Sex Offenses

Child Pornography

► Sex Offenses

Children

► Adoption

Choice Under Risk and Uncertainty

Gianna Lotito[1] and Anna Maffioletti[2]
[1]DiGSPES, University of Eastern Piedmont, Alessandria, Italy
[2]ESOMAS, University of Turin, Turin, Italy

Definition

This entry outlines what is meant by decision-making under risk and uncertainty. It illustrates the model of expected utility, its properties, and the Allais paradox as the main violation of the model. It describes the subjective expected utility model of decision under uncertainty, and the Ellsberg paradox as an example of the Knight's approach to uncertainty.

Introduction

In real economic life, many decisions are taken under risk and uncertainty, for example, investment decisions, decisions about consumption through time, buying and selling insurance, investment in new industries and countries, choosing new technologies, stock market purchases, and sales.

The literature on decision-making under risk and uncertainty can be divided in: (1) the literature concerning decision-making under risk, which includes: (a) the expected utility model (EU) and its axiomatizations (Bernoulli 1954; von Neumann and Morgenstern 1947); (b) the criticisms to the EU model (Allais 1953) and its alternatives (Kahneman and Tversky 1979; Quiggin 1982, 1993; Loomes and Sugden 1982); (2) the literature concerning decision-making under uncertainty, which can be divided into two different approaches: (a) the Bayesian approach (De Finetti 1937; Ramsey 1931; Savage 1954), according to which people assign subjective probabilities to uncertain events and these probabilities follow the rules of mathematical probability theory and (b) the approach by Knight (1921) and Keynes (1921), according to which people *cannot* assign subjective probabilities to uncertain events that follow the mathematical rules of probability theory (Ellsberg 1961; Schmeidler 1989; Tversky and Kahneman 1992).

In a situation of *risk*, one does not know with certainty the outcomes of one's choice, and the uncertainty related to a decision can be represented by an objective probability distribution over the events or states of the world which relate actions to outcomes, for example, with a gamble based on the roll of a die or a roulette wheel.

In a situation of *uncertainty,* the uncertainty relating a decision *cannot* be represented by an objective probability distribution. In this case, the individual is either considered able to attach a subjective esteem of the probability to each event (Savage 1954) – in which case decision under uncertainty reduces into decision under risk – or probabilities are not known and cannot be assigned (Knight 1921).

Choice Under Risk

Let us define now a prospect or lottery like the combination of all the possible outcomes with the probabilities of the events under which these outcomes occur.

Consider, for example, having a ticket for the following lottery L = ($100, ½;$0, ½) – where the outcomes are monetary prizes – in which a coin is tossed: in the event the coin falls Head (which occurs with probability ½), a prize of $100 is won; in the event the coin falls Tail (which also occurs with probability ½), nothing is won.

Thus, an action in a risky situation corresponds to playing a lottery or prospect, which associates each outcome to the probability of its occurrence. As a consequence, in such a situation, choice can be viewed as a choice of the preferred lottery or prospect. But how can the individual value the different prospects? One possibility is that the

individual calculates the *expected value* (EV) of the different prospects, and chooses the prospect with the highest of these values. The concept of expected value was first developed by seventeenth century mathematicians, for example, Pascal. The expected value of a lottery is the average of the monetary prizes associated with the different outcomes, weighted by their respective probabilities. The expected value of the above lottery L would be (½ * \$100 + ½ * \$0) = \$50. However, maximizing the expected monetary value does not always appear to be a satisfactory criterion to choose among different risky situations. Let us consider the two lotteries L_1 = (\$60, ½; \$0, ½) and L_2 = (\$40, ½; \$20, ½). Despite both have the same expected value of \$30, some people might prefer L_2, where there always is a probability of gaining a positive outcome, to L_1, where there is a 50% chance of winning nothing. As a further example, consider the following lotteries: L_3 = (\$100, ½; \$−1, ½) and L_4 = (\$100000, ½; \$−1000, ½). At a first sight, many people would be willing to participate to the first lottery, but would they accept so easily to play the second? It does not seem so: L_4 has a 50% probability of winning an outcome 1000 times higher than L_3, but also a 50% probability of losing an outcome 1000 times higher. However, if one calculates the expected value of the lotteries, it turns out that the EV of L_3 (\$49) is much lower than the EV of L_4 (\$49500). Clearly, expected value is not always a sufficient criterion to make a lottery attractive. An alternative is needed.

History

The first important argument on the subject was that of the mathematician Daniel Bernoulli (1738), who (independently from Gabriel Cramer) developed a new hypothesis based on the solution of a problem posed by his cousin Nicholas Bernoulli, the "St Petersburg Paradox." Suppose you have to toss a coin till "Head" comes out; the first time this happens, you stop tossing the coin and prizes are determined in the following way: if "Head" comes out at the first toss, you earn \$2; if "Head" comes out at the second toss, you earn $\$2^2$; if "Head" comes out at the third toss, you earn $\$2^3$, and so on. The probability that at every toss "Head" comes out is equal to ½, and tosses are independent from each other: then, the probability that "Head" comes out at the first toss is ½, the probability that "Head" comes out at the second toss is $(½)^2$ (that is, the probability that "Tail" comes out at the first toss times the probability that "Head" comes out at the second toss, that is, ½ * ½), the probability that "Head" comes out at the third toss is $(½)^3$, and so on. Thus, the expected value of this lottery is infinite: $2(½) + 4(½)^2 + 8(1/8) + \ldots + 2n(½)^n + \ldots = 1 + 1 + 1 + 1 + 1 + \ldots = \infty$, as the coin is thrown, if necessary, an infinite number of times and the expected prize from each toss is equal to 1. An individual who looks at the highest expected value would be willing to pay an infinite amount of money to play this lottery. But this is very unrealistic: nobody would be willing to pay more than a modest amount for it. In fact, this is a very risky lottery: it gives the opportunity to gain an increasingly bigger prize with a decreasingly smaller probability – people would not be willing to undertake this risk.

Bernoulli argued that this and similar choices could be explained by assuming that individuals do not choose the lottery with the highest *expected value*, but that with the highest *expected utility*, where the utility is represented by a function like the square root of wealth, where utility increases with wealth, but at a decreasing rate. Bernoulli introduced both the concepts of *expected utility* maximization and of decreasing marginal utility of wealth. However, Bernoulli's idea of *expected utility* maximization had little effect on the theory of decision-making under risk till the work "Theory of Games and Economic Behaviour" by von Neumann and Morgenstern was published in 1947. After that, it had soon to become the normative and positive theory of decision-making under risk.

The Expected Utility Model

The expected utility model relies on the hypothesis that the individual possesses – or acts *as if* he possesses a von Neumann-Morgenstern utility function over a set of outcomes, and when he faces alternative lotteries over these outcomes he

chooses that lottery which maximizes the expected value of this utility. In particular, von Neumann and Morgenstern start by assuming specific conditions on the preference relations between lotteries; these are necessary and sufficient to show the existence of a utility function that assigns a numerical value to the "satisfaction" of the different outcomes of the lotteries. The individual thus chooses the lottery for which the expected utility is the highest, where the expected utility of a lottery $L = (x_1, p_1; \ldots; x_n, p_n)$ is given by the sum of the products of the probability and utility over all possible outcomes, that is,

$$EU(L) = \sum_{i=1}^{n} U(x_i)p_i \qquad i = 1, \ldots, n.$$

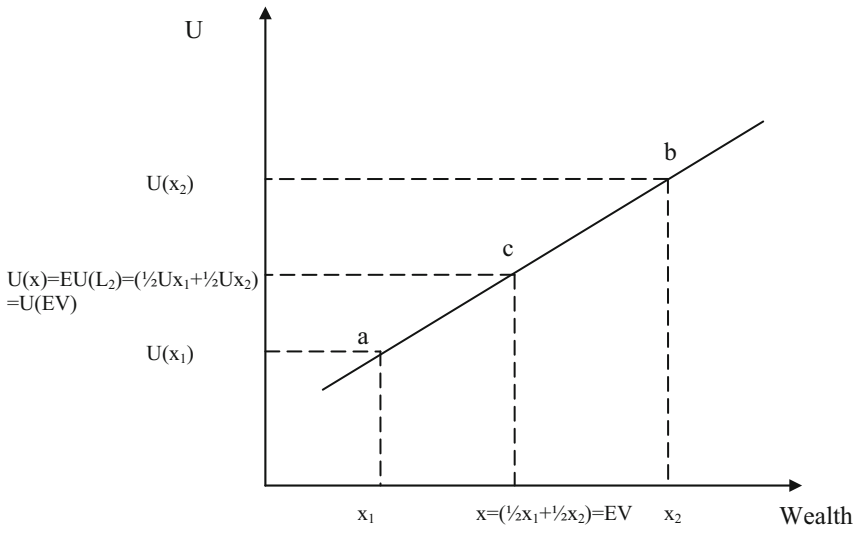
Given p_i the probability of the outcome x_i and $U(x_i)$ its utility, the expected utility of a lottery $L = (x_1, p_1; x_2, p_2)$ will be equal to $EU(L) = p_1U(x_1) + p_2U(x_2)$. The expected utility of a lottery is then the expected value of the utilities of the possible outcomes. As an example, consider the two lotteries L_1 and L_2 illustrated above. Assume that the utility function is of the form $U(x) = \sqrt{x}$, where x is the monetary value of the outcome. The expected utility of L_1 is then equal to $EU(L_1) = ½ * \sqrt{60} = \$3.87$ and $EU(L_2) = ½ * \sqrt{40} + ½ * \sqrt{20} = \$3.16 + \$2.23 = \$5.39$.

We can obtain a very important information by observing the expected utilities of these lotteries. If we compare them with the expected value (equal to \$30 for both), we notice that the expected utilities are not only different from the expected value, but are different from each other. In particular, the expected utility of the "riskier" lottery L_1 (the one which gives a 50% chance of winning nothing) is lower than that of L_2.

The von Neumann-Morgenstern utility function gives us information about the individual's attitudes toward risk.

Attitudes Toward Risk

The possible attitudes to risk of an individual can be represented graphically as follows. Consider Fig. 1 with prizes in monetary value (wealth) on the horizontal axis and their utility on the vertical axis. The straight line represents the von Neumann-Morgenstern utility function of the individual and tells us the utility he or she assigns to a given sum of money. In this case in which the shape of the utility function is a straight line, the



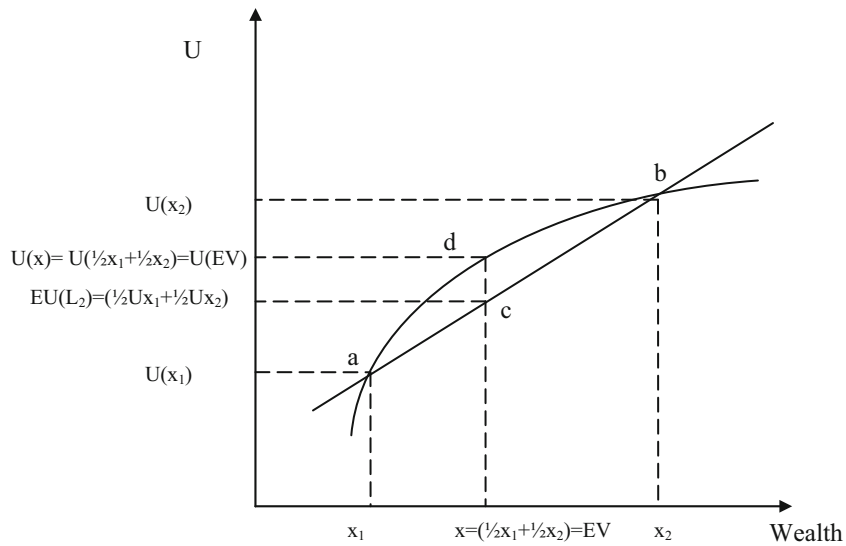
Choice Under Risk and Uncertainty, Fig. 1 von Neumann-Morgenstern utility function of a risk-neutral individual

individual is neutral to risk: when facing different lotteries he or she chooses on the basis of their expected monetary value.

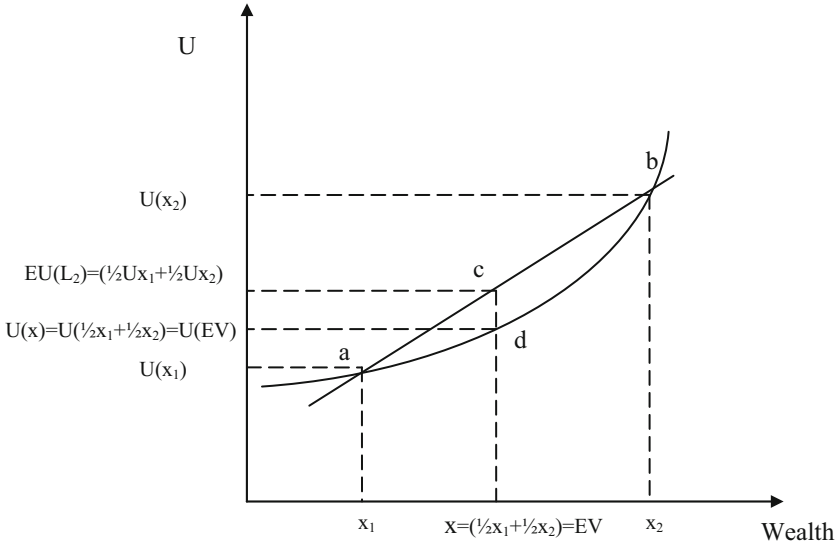
Consider the two lotteries $L_1 = x$ for certain and $L_2 = (x_1, ½; x_2, ½)$, where outcome x is also equal to the expected value of L_2. L_2 is riskier than L_1, as its prizes have a higher variability. The risk-neutral individual will not take into account this risk and will value the lotteries only on the basis of their expected value: he or she will be indifferent between them. In Fig. 1, point *a* represents the utility of the lowest outcome x_1, $U(x_1)$, and point *b* the utility of the highest outcome x_2, $U(x_2)$; point *c* is, therefore, the expected utility of lottery L_2, $EU(L_2) = ½ * U(x_1) + ½ * U(x_2)$, and is exactly midway between *a* and *b*. The expected utility of the sure outcome x, U(x) is given by point *c*. Then, an individual with such a utility function will be indifferent between having x for certain or a lottery with expected value equal to x: they have the same expected utility. The fact that L_2 is riskier than L_1 does not influence choice. For this individual thus the expected utility of a lottery is equal to the utility of a sure outcome that in turn is equal to the expected value of the lottery

Consider now Fig. 2 which represents an attitude of aversion to risk: the utility function represented here is a curve, where the utility increases with wealth, but at a decreasing rate – an individual with this utility function will not be indifferent between a sure outcome and a lottery. Suppose the individual is facing the same alternatives as before. In the graph, point *a* represents U (x_1) and point *b* $U(x_2)$; the expected utility of lottery L_2 will be the weighted average of these two utilities and will be given by point *c*, in the middle of the segment unifying *a* and *b*, ½ $U(x_1)$ + ½ $U(x_2)$. However, we see now that the expected utility of the sure prize x is given by point *d*, which is above *c*: $U(x) > ½ U(x_1) + ½ U(x_2)$. For this individual thus the expected utility of a lottery is lower than the utility of a sure outcome that is equal to the expected value of the lottery. The individual who is averse to risk will choose the sure alternative to a lottery which gives an expected prize equal to the sure one – he or she prefers the expected value of the lottery to the lottery itself. The individual is averse to the risk connected to the lottery.

We should note that this risk aversion is implicit in the shape of the utility function, which is concave. As an illustration of this, consider lotteries (\$60, ½; \$0, ½) and (\$40, ½; \$20, ½) above. In calculating their expected utilities, we have assumed a form for the utility function equal to $U(x) = \sqrt{x}$, which has a shape like in



Choice Under Risk and Uncertainty, Fig. 2 von Neumann-Morgenstern utility function of a risk-averse individual



Choice Under Risk and Uncertainty, Fig. 3 von Neumann-Morgenstern utility function of a risk-loving individual

Fig. 2. This is a utility function implying risk aversion, as shown by the fact that when calculating the expected utilities of the above lotteries, we find that the riskier lottery has a lower expected utility than the "safer" lottery.

The case of the utility function of an individual with a proneness to risk is given in Fig. 3. It can be seen here that for such an individual the expected utility of a lottery is higher than the utility of a sure outcome that is equal to the expected value of the lottery. The individual who is risk loving will choose the lottery which gives an expected prize equal to the sure outcome instead of the sure alternative – he or she prefers the lottery to the expected value of the lottery itself. This individual loves the risk connected to the lottery.

The individual preferences toward risk can also be expressed using the concepts of (1) *certainty equivalent* (CE) and (2) *risk premium* π.

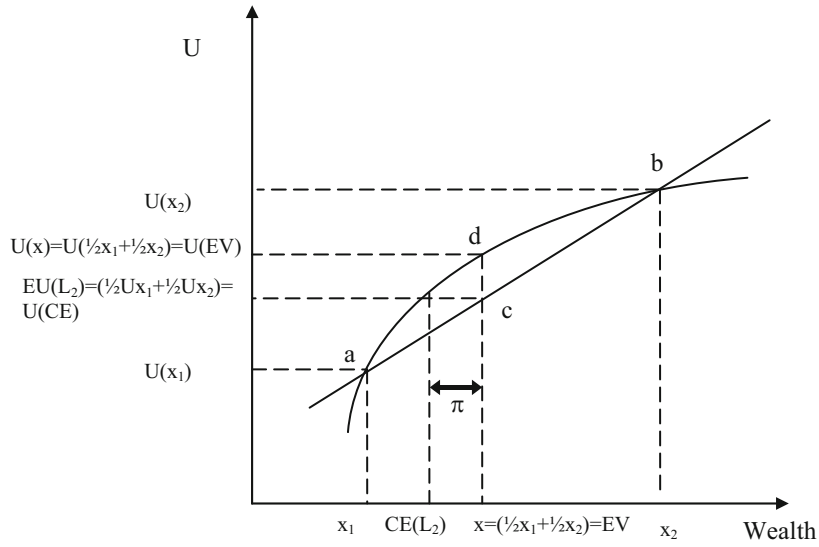
1. As an example, consider the lottery L = ($100, ½; $0, ½). The expected value of this lottery is 50. Consider asking this individual what is the amount of money which makes her indifferent to the lottery. This is defined as the *certainty equivalent* of the lottery. Therefore, the utility of the CE of the lottery is defined as equal to its expected utility – U(CE) = EU(L).

   Suppose the individual is indifferent between the lottery and $40 for certain, that is, his CE for the lottery is equal to 40, less than the expected value of the lottery. It follows that the subject would accept any sum >40 instead of the lottery and the lottery to each sum <40. In this case, we say that the subject is averse to risk – U(CE) < EV(L). This is represented in Fig. 4. The reverse inequality defines risk proneness, and the equality defines risk neutrality.
2. The risk premium π is defined as *the maximum part of the expected value that the subject is willing to give up to avoid the risk associated to the lottery*, π = EV(L) – CE (see Fig. 4).

Consider the example of a CE for the above lottery equal to 40; this means that the individual is willing to give up at most $10 of the expected monetary value of the lottery, that is, the risk premium is 10. The sign of the risk premium gives the attitude to risk for that lottery. In general, when for any lottery π is positive, the individual is averse to risk; when π is negative, he is risk lover; when π = 0, the individual is risk neutral.

The most common analytical measures of risk aversion have been introduced by Arrow (1964) and Pratt (1964) and extensively used in finance, insurance markets, health, and game theory.

Choice Under Risk and Uncertainty,
Fig. 4 Utility function of a risk-averse individual with certainty equivalent and risk premium measures



The Allais Paradox

The most important implication of the specific form of the expected utility preference function $EU(L) = \sum_{i=1}^{n} U(x_i)p_i$ is linearity in the probabilities (Marschak 1950; Machina 1987; Raiffa 1968). For this reason, most of the empirical investigation of the expected utility hypothesis has focused on this property (MacCrimmon and Larsson 1979; Allais and Hagen 1979; Starmer and Sugden 1989; Starmer 2000), revealing widespread systematic violations. The best-known violation of linearity in the probability is the Allais paradox (Allais 1953). Consider that an individual is asked to make a pairwise choice between the two following pairs of lotteries (outcomes in French francs as in Allais 1953, p. 527):

- A 100% chance of 100 millions
- B { 10% chance of $500 millions; 89% chance of $100 millions; 1% chance of 0
- C { 11% chance of $100 millions; 89% chance of 0

or

- D { 10% chance of 500 millions; 90% chance of 0

If the individual has expected utility preferences, a choice of A over B in the first pair implies a choice of C over D in the second pair. However, empirical findings show that the modal choice is for A over B in the first couple of lotteries, but D over C in the second couple. This pattern of choice violates expected utility. In fact, A preferred to B implies U(100 millions) > .10U (500 millions) + .89 U(100 millions) + .01U(0), that is, .11U(100 millions) + .89U(0) > .10U (500 millions) + .90U(0), while D preferred to C implies .10U(500 millions) + .90U(0) > .11U (100 millions) + .89U(0), which is a contradiction.

The widespread and systematic empirical violations of the linearity property has put under discussion the descriptive validity of the theory, giving rise to a growing body of literature of new theoretical models of choice under risk as, for instance, Prospect Theory (Kahneman and Tversky 1979; Tversky and Kahneman 1992), Rank Dependent Utility Theory (Quiggin 1982, 1993), Regret Theory (Loomes and Sugden 1982).

Choice Under Uncertainty

The first distinction between choice under risk and choice under uncertainty was made by Knight (1921) and by Keynes (1921). They define a choice under uncertainty when we deal with a choice in which the probability of the occurrence of an outcome is not known. Savage (1954)

extended the theory of Expected Utility to uncertainty and called it Subjective Expected Utility. According to Savage, all individuals can have a subjective esteem of the probability attached to an event. If this is true, then uncertainty is reduced to risk. Moreover, according to Savage, this individual estimate of probability follows the mathematical rules of probability. Consider the following lottery: "you will receive $100 if tomorrow at noon the temperature in Rome is higher than 30 degree Celsius and $0 otherwise." The probability that we assign to the event "at noon the temperature in Rome is higher than 30 degree Celsius" and the probability that we assign to the opposite event "at noon the temperature in Rome is 30 degree Celsius or less" follow the rules of probability, that is, they sum to one. In addition, our probability estimate should be inferred from our choice between lotteries.

Ellsberg (1961) pointed out that this is not true, and hence uncertainty cannot be reduced to risk. Consider the following example given by Ellsberg. You have in front of you two urns: Urn I and Urn II. Both urns are opaque so you cannot see inside. You are told that in Urn I there are 100 balls: 50 are black and 50 are red. Instead, Urn II contains 100 balls that can be either black or red but you do not know in which proportion.

You are asked to bet on a color and choose the urn from which to draw the ball simultaneously. If the ball that you have drawn is of the same color you have chosen, then you win $100, otherwise you will get nothing.

According to Ellsberg, most of the people will decide to draw the ball from Urn I, which contains 50 red and 50 black balls, independently from which color they have chosen to bet on. In this case, in fact the probabilities are known to be 50/100 for Red and 50/100 for Black.

The proportion of the red and black balls in the second urn is not known and so we do not know which probability to assign to red and black. Avoiding to choose to draw a ball from Urn II is called by Ellsberg *uncertainty aversion*.

According to the theory of Savage, we should be indifferent between the two urns since the probability of getting red can be represented by a second order probability distribution (a probability distribution over probability), whose expected probability is 50/100 as in the case of Urn I. However, people prefer to bet on the first urn showing that they prefer to draw a ball from an urn in which the probabilities of the two color balls are precisely defined.

Aversion to uncertainty creates a problem to Savage theory also for another reason. Let's consider again the two urns. If you always choose to draw a ball from Urn I, whatever color you prefer, this implies that you consider the probability of getting red plus the probability of getting black in Urn I different from the probability of getting red plus the probability of getting black in Urn II. If this is the case, then, the sum of your probability estimates of the two color balls for the two urns is going to be different from 1. In particular, ambiguity aversion implies that your estimate of the probability of red plus your estimate of the probability of black in Urn II is less than 1. Your probability estimates thus are not following the mathematical rules of Savage theory. Intuitively, the difference of the sum of the two probabilities from one is the room that we leave to the existence of uncertainty, that is to say, the possibility of occurrence of some unexpected situation. As a consequence, it is very difficult to deduce our probability estimate from our choice as pointed out by Ellsberg. Ellsberg's work has been confirmed by a substantial number of empirical and theoretical research contributions (Camerer and Weber 1992; Trautmann et al. 2008; Machina and Siniscalchi 2014; Gilboa and Marinacci 2016).

Cross-References

▶ Risk Management, Optimal

References

Allais M (1953) Fondements d'une théorie positive des choix comportant un risqué et critique des postulates et axioms de l'école Américaine. Econometrica 21(4):503–546

Allais M, Hagen O (eds) (1979) Expected utility hypothesis and the Allais paradox. D. Reidel, Dordrecht

Arrow K (1964) The role of securities in the optimal allocation of risk-bearing. Rev Econ Stud 31:91–96

Bernoulli D (1954) Specimen theoriae novae de mensura sortis. (Commentarii Academiae Scientiarum Imperialis Petropolitanae, 1738). Econometrica 22:23–26. (Trans. as Exposition of a new theory on the measurement of risk)
Camerer CE, Weber M (1992) Recent developments in modeling preferences: uncertainty and ambiguity. J Risk Uncertain 5(4):325–370
de Finetti B (1937) La prévision: ses lois logiques, ses sources subjectives. Ann l H Poincaré 7(1):1–68
Ellsberg D (1961) Risk ambiguity and the Savage axioms. Q J Econ 75:643–669
Gilboa I, Marinacci M (2016) Ambiguity and the Bayesian paradigm. In: Arló-Costa H, Hendricks VF, van Benthem J (eds) Readings in formal epistemology. Springer, New York
Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:273–291
Keynes J,M (1921) A treatise on probability. McMillan, London
Knight F (1921) Risk, uncertainty and profit. Houghton-Mifflin, Boston
Loomes G, Sugden R (1982) Regret theory: an alternative theory of rational choice under uncertainty. Econ J 92:805–824
MacCrimmon KR, Larsson S (1979) Utility theory: axioms versus 'paradoxes'. In: Allais M, Hagen O (eds) Expected utility hypothesis and the Allais paradox. D. Reidel, Dordrecht
Machina M (1987) Choice under uncertainty: problems solved and unsolved. J Econ Perspect 1(1):121–154
Machina M, Siniscalchi M (2014) Ambiguity and ambiguity aversion. In: Machina M, Viscusi K (eds) The handbook of the economics of risk and uncertainty. North-Holland, Oxford
Marschak J (1950) Rational behavior, uncertain prospects, and measurable utility. Econometrica 18:111–141
Pratt JW (1964) Risk aversion in the small and in the large. Econometrica 32:122–136
Quiggin J (1982) A theory of anticipated utility. J Econ Behav Organ 3:323–343
Quiggin J (1993) Generalized expected utility theory. Kluwer, Dordrecht
Raiffa H (1968) Decision analysis: introductory lectures on choices under uncertainty. Addison-Wesley, Reading
Ramsey FP (1931) Truth and probability. In: The foundations of mathematics and other logical essays. Harcourt, Brace, New York
Savage L (1954) The foundations of Statistics. Wiley, New York
Schmeidler D (1989) Subjective probability and expected utility without additivity. Econometrica 57(3):571–587
Starmer C (2000) Developments in non-expected utility theory. J Econ Lit 38:332–382
Starmer C, Sugden R (1989) Violations of the independence Axiom in common ratio problems: an experimental test of some competing hypotheses. Ann Oper Res 19(1):79–102
Trautmann ST, Vieider FM, Wakker PP (2008) Causes of ambiguity aversion: known versus unknown preferences. J Risk Uncertain 36(3):225–243
Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertain 5:297–323
von Neumann J, Morgenstern O (1947) Theory of games and economic behaviour, 2nd edn. Princeton University Press, Princeton

Civil Law System

Pier Giuseppe Monateri
SciencesPo, Ecole de droit, University of Torino, Torino, Italy

Definition

This entry defines what is meant today for civil law; sketches an outline of its historical background; individualizes the two main models of it, German and French; and points to the difficulties of comparison with the common law, as to the hardships of a work of harmonization.

Introduction: What Is Civil Law as a Legal Family?

What we call "civil law system" is indeed a family of different legal systems tracing their historical roots to the Roman law.

As such, this family of legal systems is differentiated, today, especially in regard to the other two major legal families existing in contemporary world legal landscape: the common law legal family and the Sharia of the Islamic legal model. As well as the civil law, the "common law" is a set of highly differentiated systems of law sharing the same origin to be found in the history and development of the English common law. Differently the Muslim Sharia is supposed to be a unique system of principles and rules, based on the divine revelation contained in the Koran, even if its interpretation may vary very greatly in different jurisdictions, cohabiting, also, with European-like codes and modern constitutions, and today is, on the average, applied only to the *status personae*, the personal condition of the subject, as marriage, divorce, inheritance, and other related matters (Samuel 2014).

This given, it is manifest that when we speak of common and civil law, as the two major variants of the Western legal tradition, we make reference to the different *legal origins* of modern systems, implying that these differences are still molding the actual structure of our laws (World Bank 2003).

Historical Background of the Civil Law Origins

The term "civil law" is an English term used to translate the *jus civile* or the proper Roman law as it evolved from classical times to the end of the empire when it became codified by Justinian, from 529 to 534 AD, in his codes, constituting an ordered collection of a mass of writing known as the *Corpus Juris Civilis* or *The Body of Civil Law*. The work was planned to be divided into three parts: the *Code* as a compilation of imperial enactments, the *Digest* or *Pandects* composed of *advices* given by older Roman jurists on different points of the law and deemed to have authority for their learned character, and finally the *Institutes* conceived as a textbook for law students at the newly established law school of the empire in Beyrouth.

Tribonian has been the editor in chief of this massive work, thought to represent the whole of the jurisprudential tradition evolved from early Roman times up to the date of the compilation.

It is important to note two main facts:

First, it is the fact that the Roman Empire at that time was split into two parts and that this compilation was enacted, having force of law, only in the eastern part of the empire speaking Greek. In this way, the Justinian compilation, quite exotically, has been written in Latin for an empire speaking Greek and was never enacted as such in the West, but influenced its legal progress in the strongest possible way, something which defeats any of our actual understandings of the working of law.

Second, this enterprise has marked a total revolution of Roman law, changing completely its style and its structure. Roughly speaking, classical Roman law was an *oral* law, without codes, but only with pieces of legislation passed by the various political assemblies. There was *not* a formal system of legal education, each one having to learn the law from a practicing lawyer, and especially there were not regular courts of law (Glenn 2000).

The Roman magistrate directing the trial, the praetor, was a politician, appointed for 1 year, controlling only the *form* of actions pleaded before him by the parties. Then, to afford the trial, he had to nominate a *judex*, a "judge," a layman, to be agreed by the parties. In this way he was more an *arbitrator* than a *judge*. Just for this reason, the learned opinion of jurists of great reputation played such an important role: they had to advice the *praetor* and the *judex*, as laymen, upon difficult and disputed points of the law. Moreover, classical Roman law was ruling *only* Roman citizens, namely, only male adults, whose father was already dead and belonging to Roman families, a very small proportion of the inhabitants of the empire. Roman law has never been the clue of the empire: Egypt was ruled by Egyptian law, Greek cities by their own laws, and so on. Only in 212 AD emperor Caracalla extended, for fiscal reasons, the citizenship to all the inhabitants of the empire.

This "classical model" evolved, then, slightly over time into the opposite one, which was finally molded by Justinian, having a central court of justice at the imperial chancellery, a formal legal education at the law school in Beyrouth, and a fixed system of written sources collected into the *Corpus Juris* and universally applicable to the whole of the empire. By this fact, we can say that the final shape of Roman law, left in inheritance to the Middle Ages, was exactly the opposite of its beginnings: from an *oral* law, administered by laymen, valid only for the very few, to a written law, administered by professionals, universally valid. It is important to remember that all this happened in the East and *not* in the western part of the empire which remained a patchwork of different laws: old Roman law, canon law, and the various laws of

the "German" nations, Goths, Franks, and others, which occupied the West (Berman 1983).

This *eastern* legacy became, anyway, extremely important in the West for theological political reasons linked to the birth and development of a renewed Western Sacred Empire from Charlemagne, 800 AD, to the establishment of the first modern university in Bologna (1174 AD) and on.

The "great space" of continental Europe became to be shaped in "catholic" terms: the Sacred Empire was to be thought as a single "body," because eating the same Holy Communion, all his inhabitants shared the same flesh. The compilation made by Justinian became to be regarded as a real "revelation" of the law for all mundane affairs not strictly confined to the church or to be left to morality. Indeed, this compilation was the only extant remain of the law, because it was written in bounded volumes of parchments, made to last, whereas all the previous scripts were on papyrus paper, necessitating to be regularly copied to be preserved, and so went quite completely lost in the barbarian west. Besides, it was much more comprehensive and well ordered than any existing barbarian compilation of laws.

In this way, nobody really enacted the Justinian compilation as positive law in the West, but it was thought to be the *ratio scripta*, the codified reason, of the law of a sacred unitary political body ontologically grounded on the Holy Communion of all its inhabitants.

This sacred, and universal, as well as *rational* character of the compilation explains why it became the basis of the university teaching of the law at Bologna, the first university established in the West, from which sprang Padua, Paris, Oxford, and Cambridge, where indeed Roman and *not* English law was taught. But the English Kingdom always refused to become a *terra imperialis* and so always refused to give any practical application to Roman Justinian law. On the continent, this common teaching shaped similarly, all over the places, the legal mind of professionals, and it was deemed applicable, as a law of reason and last resort, in all cases not patently covered by local legislation.

This legal landscape formed the era of *jus commune* in continental Europe to be broken only by the advent of modern codifications at the end of the eighteenth century. This also explains, in comparison with the English legal system, the highly intellectual character of civil law: it was a university scholarly law. Besides, on the continent, the use of writing never went completely abandoned as it almost happened in England. English *jury* trial, as an *oral* pleading, was quite a necessity given the incapacity of the jurors to read documents, whereas the continent could adopt a more sophisticated system of trial, based on documents and administered by clerks (Watson 2001).

Law and Modern Codifications: The French Model

As we have seen in the previous paragraph, continental law evolved as a *jus commune* of a common empire, based on a theory of the Justinian compilation both as sacred and as rational. Of course, the destiny of this political theological complex was to come to an end with the growing antagonism of France, Spain, and Germany, and especially with the 30-year war (1618–1648) of religion following the protestant reform.

It is out from this war that emerged on the continent the idea of the modern sovereign state.

The inter-Christian war was not terminable *but* in pure political terms: a sovereign absolute on his territories deciding also the faith of his subjects. This rising of the local princes to the status of absolute independent rulers fractured the catholic space of the empire into different territories with different jurisdictions giving rise, with the peace treaty of Westphalia (1648), to the modern system of interstate relationships known as international law. Each new sovereign became like a local, territorial bound, piece of the fractured mirror of the global universal authority of the empire, which was reflecting God's government of the world.

It is quite natural, then, that from a concept of the sovereign, as an absolute concentration of local political power, emerged the idea that it was in the hand of this sovereign to ordain and establish the laws of his realm; and since the imagery linked to Justinian was still that of him

as the template of the lawgiver, the various monarchies tried to follow his model in projecting codes of a comprehensive, universal, and rational character for their own domains (David and Brierly 1968).

The first project was that of Frederick I of Prussia, then performed by Frederick II, leading to a *Project eines Corporis Juris Fridericiani* (1749–1751), drafted by Samuel von Cocceji. The same name of the project is displaying the Justinian ambitions of these modern sovereigns. This project led to the so-called Allgemeines Landrecht or the general laws for the Prussian states finally codified in 1794 under the supervision of Svarez and Klein, who were under the orders of Frederick the Great. This project is of extreme importance since it represents the idea that the sovereign state can shape society at it wishes and that he has not only the political power of war and peace but also that of ordering society by legislation. In this way Justinian law which was really a universal legislation served as a template for local legislations of the modern states, breaking the previously prevailing universal conception of space.

Following this German example, Maria Theresa, Empress of Austria, decided, about 1770, to charge a committee with the task of preparing a code of all her lands. After 40 years of preparatory works directed by Karl Anton Freiherr von Martini and Franz von Zeiller, this project was enacted in 1811 as the *Allgemeines bürgerliches Gesetzbuch* (ABGB), the Civil Code of the Austrian-Hungarian Empire.

What happened in between was one of the real major breaks in all European political history: the French Revolution. From a legal point of view, the revolution captured the sovereign within the state, making him no longer the possessor of the state but one if his constitutional organs, and finally sentenced the king to death for high treason, conferring an all-mighty power to the popularly elected legislative assembly. The revolutionary government went on performing a complete subversion of the existing law, hooting almost the 75% of judges, dissolving the Bar, and closing all the law schools. The new faculties of law were founded, the legal profession was completely reorganized, and a new judiciary was established inventing the modern pyramid of courts we can find in every civil law jurisdiction. It is made of many tribunals, in quite every district, to judge on cases of first instance, then of fewer appellate courts to review their judgments, and finally of one *Cour de Cassation* established to grant a uniform application of the law.

Meanwhile, many measures were adopted to grant a legislative unity of the state, and at the end of the revolution, when Napoléon I became emperor, on 21 March 1804, he installed a commission to draft a code and, on the same year, he enacted the French *Code Civil* or *Code Napoléon*, officially the *Code civil des Français*, as a real *liberal constitution* of the civil society.

The whole apparatus to reach this goal was once again derived from Roman templates. After all the revolution was conceived to reestablish a kind of "Roman Republic," giving back to the people all the powers and prerogatives usurped by the kings and the church; and the first title assumed by Napoléon himself was that of First Consul of this polity.

He participated to the most of the discussions in the committee and imposed a literary style to "his" code inspired by the principles of brevity and clarity, as it was thought to be a code for the commons and not for the specialists. This same code became to be surrounded by a constellation of other codes: the penal code, the code of civil procedure, the code of commerce, and the code of criminal instruction. The civil code was divided into three parts: persons, property, and "the different ways of acquiring and transferring property," a section mainly devoted to contracts, torts, and unjust enrichment. The code is very liberal considering marriage as a contract, defending property as an absolute right, shaping contract as an agreement based on the free choices of the parties, and considering negligence as the basis of any liability.

In this way, France became the real model of any modern codified system, and her codes had an immense impact on the other countries from Italy, Poland, Spain, and Greece, to Latin American legal systems, then to Egypt, Syria, and many other systems in Africa and in Asia.

So to speak, France *is* what we have in mind today when we speak of a civil law jurisdiction.

Its main features are codes covering the whole of the legal field and a judiciary diffused all over the country and organized on the three levels of tribunals, courts of appeal, and a central court of cassation.

It is important, here, to underline the pivotal role assumed by legislation confiding to it the power to order society in all its details, because of its revolutionary political role. The center of gravity of the revolution has been the legislative assembly, and the revolution was mainly a revolution of laws, collapsing all the structures of the Ancien Régime, something which never happened in England, where this ideology of legislation was rejected also by liberals like Edmund Burke, in favor of a "sublime" conception of an oral law and an unwritten constitution as instantiated in judicial decisions, remembering, anyway, the extremely *elite* nature of the English judiciary having only *one* High Court in London, with an appellate division, submitted to the nine justices of the House of Lords (now called the Supreme Court of the United Kingdom). The French arrangement of the judiciary is extremely more diffused: the English Law Lords are nine deciding approximately 60 cases per year; at the Court of Cassation, we find more than 150 judges deciding quite 7,000 cases a year (Milo et al. 2014).

The most important point is anyway that legislation, and the rational constructivist idea of the possibility for it to *design* society, lies at the basis of the French legal system molding also the French legal style. Courts are rendering very brief decisions adopting the same style of the code, almost one page long only, whereas an English or American decision can be also 40 or 50 pages long, reporting not only the impersonal view of the court, as a unanimous organ, but all the opinions of minority and majority justices.

There is finally another factor to be remembered which is normally underscored. Parallel to the general jurisdiction, the French system adopted a special administrative jurisdiction, confined to cases involving the public administration, having its *apex* in a peculiar French institution: the *Conseil d'État*. The very existence of this institution was singled out by authors like Dicey as the major difference between the English and the French system. In this way, the common law idea of judicial review of administrative acts is not followed in France. Normal judges have no jurisdiction over state acts: these can be questioned only behind the administrative jurisdiction and the Conseil d'État, an organ which is not only working as a court but also as a counselor of the administration in producing by-rules and acts. Under this respect, no two other systems could be more different.

Civil Law and Modern Codifications: The Rise of the German Model

As we have seen, Prussia had a code before France, but then the Napoleonic Empire extended French domination all over Europe, transplanting French patterns and methods all across the continent, up to when the French Army was defeated in Russia in 1812. The Germans lived the time between 1812 and the final defeat of Napoléon at Waterloo as an era of *national wars of liberation* against the French. After the Vienna Congress of 1815, Germany was restored but as a constellation of 39 different sovereign states: Prussia, Schleswig-Holstein, Bavaria, and so on. Anyway, its "space" (Reich) was deemed unitary from the standpoint of sharing a common culture, a common language, *and* a common university teaching. So attempts were made for having also a common legislation overpassing the differences between the various states notwithstanding the lack of a political unity (Wieacker 2003).

Thibaut was an author who sponsored the theory of adopting a German version of the French code. His idea was rejected by the most prominent German law scholar of all times Friedrich von Savigny. In an outstanding article (*Vom Beruf unserer Zeit fuer die Gesetzgebung und Rechtswissenschaft*), he traced a parallel between law and language (likely to be derived from the Scottish Enlightenment) in order to block the adoption of a foreign legislation. As the language is a complex spontaneous order, so it is the law.

Law and language are evolving orders that no single group of human minds have consciously designed nor can control. They are decentrated orders, like markets (Hayek). So it is impossible and hazardous for legislation, as a consciously designed order, to try to mold the whole of society. Society is different from the state, which is one of the many purposive organizations pursuing their goals within society. It follows that the overall order of society cannot be designed, but can only evolve piecemeal.

This theory is rather understandable if we remind that there wasn't a unitary state in Germany, so that effectively there was no possibility for a central authority to mold the law, nor there was any unitary judiciary to promote it. What was unitary in the various German states was the university system. A student could also spend a term in Munich and the next term in Berlin; and what Savigny proposed, after the feelings raised by the very conception of the wars of liberation to build up a newer Germany, was to entrust the development of the law to the legal science (*Rechtswissenschaft*) as practiced by the German *professoriat.*

If law is like language and language is a depository of culture, it makes no sense to adopt a foreign law and destroy our culture while engaging in liberation and the making of renewed Germany. Law and language lie in the *spirit of the people* (*Volksgeist*). Only a scholar can have a good insight over it, because of his learning, to be able to produce a well-conceived framework of concepts to give it voice, creating a kind of scholarly made law (*Juristenrecht*) different from both judicial-made law and from legislation. And, after all, Germany was to be considered as the real heir of the "space" of the empire (*Reich*), and as such went on, and was going on, elaborating the *jus commune*, the actualized version of the Roman law. This law was not a piece of ancient history in Germany but an *actual* system of living law. In this way, Roman law was no more an alien system, but it really became, in many centuries, part of the national spirit. Indeed, Savigny's major work was entitled *Der System des heutigen Roemischen Rechts*, "The System of the Actual Roman Law."

Here, we may find a version of the civil law totally opposite to that of the French. Where France claims to be "republican," but she is indeed the continuation of the imperial model of Justinian, entrusting law to legislation, with the possibility of a political design of society, here, Germany is representing the ideal of the "classical" Roman law as a law practically *without* legislation, and certainly without codes, slightly evolving through learning, as the great jurists of Rome did *before* Justinian and as the great lawyers of the *jus commune* did after Bologna. France is claiming a continuity with Roman templates of codification, but Savigny is claiming a deeper and strong continuity where legislation is but an episode of a much more complex story of the civil law tradition.

If we perceive this, we can easily spot how codes are an unnecessary feature of a civil law system and maybe are contrary to its original nature.

Savigny prevailed against Thibaut and Germany went on developing "scientifically" the Roman law. But when, with the war of 1866 against Austria and of 1870 against France, Germany was unified in the form of the Second Empire, the pressure for having a common legislation became too strong. This pressure could anyway be filtered by the already established institution of the *professoriat* as a real factor of the legal progress. So professors started to work on the idea of making a new code different from the French one and based on the "concepts" used to elaborate their own actualized version of Roman law (*Begriffsjurisprudenz*), especially Windscheid, a well-known author of one of the major textbooks on the *Pandects* paved by his scholarship, the way to a first draft of the code in 1888. A committee of 22 members, comprising not only jurists but also representatives of financial interests and of the various ideological currents of the time, compiled a second draft. After significant revisions, the BGB (*Burgerliches Gesetzbuch*, Civil Code) was passed by the Reichstag in 1896. Political authorities gave 4 years to the legal profession to study and learn the new legislation, which was put into effect on 1 January 1900 and has been the central codification of Germany's civil law ever since.

The BGB served as a template for several other civil law jurisdictions, including Portugal, Estonia, Latvia, Japan, Brazil, and Greece. It never had, anyway, the same world impact as the French code. What had a tremendous impact all over the civil law countries were German scholarship and the German method strongly influencing Italy, Spain, Latin America, and quite all the jurisdictions that maintained a French-like legislation (Merryman and Pérez-Perdomo 2007).

So, after all, also Germany became a codified system, and quite all civil law jurisdictions can be deemed to be a "hybrid" of French legislation and German scholarship.

What is peculiar is that the two codes, French and German, are really very different. The German code, especially, possesses a General Part (A*llgemeiner Teil*), which does not exist in the French code. In this General Part, we can find all the general concepts to be adopted to grasp the specific parts devoted to contracts, torts, and property. This different approach is obviously indebted to the fact that this code has been elaborated by professors and that they have been able to act as a unitary factor to reach a national goal.

Anyway, the Germans structured the judiciary in quite the same French way and maintained a separate administrative jurisdiction as in France.

Conclusion: The Problems of Harmonization and of Comparison Between Common and Civil Law Jurisdictions

All this, the mixing of the French and German patterns, is giving to civil law, considered as a general tradition, her intellectualistic flavor as well as her pro-legislation biased aspect.

When we speak of civil law jurisdictions, we mean systems that (1) have codes, (2) have a similar and diffused judiciary handling many more cases than a common law jurisdiction, (3) possess a separate – seemingly pro-state biased – administrative jurisdiction, and (4) know a much stronger and active role in the legal development of scholars and universities (van Caenegem 2001).

Notwithstanding this general image of the civil law, there are some myths to deconstruct about the comparison of civil and common law systems. First of all one is the myth that civil law *is* legislation and common law *is* a judge-made law.

Today, the most of legal matters in common law countries are covered by statutory law. Corporate governance, for instance, is always legislative also in these countries, as it is sale of goods or secured transactions. On the other side, it is true that the legislation of the continental codes is very broadly conceived, so that the role of judges in developing the sense of the codes cannot be underestimated. Case law is as important to understand a provision of a civil code as it is to know what the common law is on a certain point.

Second, it is not true that legislation is a permanent and overwhelming factor in civil law countries. They lived for centuries without codification, and we may find, as in the case of Savigny, theories of the essence of the civil law which are directly antagonistic to the role of legislation.

Third, it is true that the civil law appears more "conceptualized," for the role always played by universities in her elaboration, but we cannot overpass the role of theory in the United States. It would be hard to consider American law without considering that each case is based upon a *doctrine* and that it is much more American scholarship, than state case law, to give a picture and a frame of what this law is and to influence the rest of the world, and we cannot bypass the role of great law schools in the practical organization of the elite of the legal profession, their ways of thinking, of elaborating solutions, and so on. From a civilian perspective, an American piece of legal scholarship is much more based on theory than it is, today, an average civil law writing displaying more erudition and knowledge than intellectual claims.

It is rather to be accepted that both families are a different compound of different factors always acting, sometimes in competitive ways, in the legal history: legislation, judicial decisions, and scholarly writings. The different mixtures of these elements are marking the difference between France and England, but it is marking the difference between England and the United States, also,

as it marks a difference between France and Germany (Ginsburg et al. 2014).

What is really different in common and civil law is the figure of the judge and the fact of having a separate administrative jurisdiction.

Judges in common law are fewer and decide a much lesser number of cases. This is something in search for an explanation. There are approximately 6,000 judges in France and 600 judges in England. Besides, a common law judge is an old member of the Bar (the United Kingdom), or she is directly appointed by the political power at state or federal level (the United States). A civil law judge is the winner of a public competition for recruitment. It means that you become judge when you are young, just maybe practicing the law for a few years, and then you make a *judicial career* from the last of tribunals to the chair of the president of the Court of Cassation, whereas there is scarcely something as a judicial career in the United States, so few being the case of persons appointed as state or federal circuit judges then becoming appointed at the Supreme Court. Under this respect, the two systems cannot be more divergent. This factor depends heavily on the costs of justice. Civil law is cheaper, and that's also why it is normally longer; but no serious attempt has been made to understand precisely why, and this certainly does not depend on Roman origins.

The fact of having a separate administrative jurisdiction is also of extreme relevance. This fact, again, cannot be traced back to the Roman origins of the civil law systems; rather, it is a by-product of political modernity: the rise of an absolute state on the continent and the absence of a political upheaval similar to French revolution in the common law world.

It is strange to note the following paradox: in common law, ordinary jurisdiction is much more politicized in the sense that the judge can be appointed directly by the political power, but the civil law is granting more room for state action by creating an administrative compartment separated from ordinary jurisdiction.

But is the separation of ordinary and administrative jurisdictions connaturate to a civil law tradition? One could really wonder. For centuries, again, there was not such a separation, and it is much more likely to be due to the *form* assumed by political power on the continent of Europe than to deep legal structures linked with distant origins.

Finally, what is certainly absolutely distant, even today, is the *style* of these two families of laws. There is scarcely any similitude between a French and an American judicial decision, as there is not a common way to handle precedents, and also the modes of interpreting statutes are rather distant. In a sentence we could say that the apparently politically flat world of globalization is still *striped*, fractured, and discontinued by the *legal styles* (Zweigert and Koetz 1998).

To what extent, if any, these legal styles have an economic impact is a question open to investigation. What it certainly represents is a legal *duality* of the West, and especially of Europe, displaying two different appearances of what we call justice, rendering any work for harmonization harder than expected.

Summary/Conclusion/Future Directions

There are two different main versions of the civil law, German and French, as there are similarities and differences between civil and common law which are hard to grasp. The major difference lying in the different *styles* of these legal traditions hampering any actual conscious work of harmonization as they represent complex spontaneous orders which can but imperfectly been managed by purposive design.

Cross-References

- ▶ Administrative Law
- ▶ Globalization
- ▶ Law and Economics, History of

References

Berman HJ (1983) Law and revolution: the formation of the western legal tradition. Harvard University Press, Cambridge, MA

David R, Brierly JE (1968) Major legal systems in the world: an introduction to the comparative study of law. The Free-Press/Collier Macmillian, London
Ginsburg T, Monateri PG, Parisi F (2014) Classics in comparative law. Edward Elgar, Cheltenham/Northampton
Glenn HP (2000) Legal traditions of the world: sustainable diversity in law. Oxford University Press, Oxford
Merryman J, Pérez-Perdomo R (2007) The civil law tradition: an introduction to the legal systems of Europe and Latin America. Stanford University Press, Palo Alto
Milo JM, Lokin JHA, Smits JM (2014) Tradition, codification and unification: comparative historical essays on developments in civil law. Intersentia Ltd, Cambridge, UK
Samuel G (2014) An introduction to comparative law theory and method. Hart Publishing, Oxford/Portland
van Caenegem RC (2001) European law in the past and the future: unity and diversity over two millennia. Cambridge University Press, Cambridge, UK
Watson A (2001) The evolution of Western private law. John Hopkins University Press, Baltimore
Wieacker F (2003) A history of private law in Europe with particular reference to Germany. Clarendon, Oxford
World Bank (2003) Doing business in 2004. Understanding regulation. The World Bank, Washington, DC
Zweigert K, Koetz H (1998) Introduction to comparative law. Oxford University Press, Oxford

Class Action and Aggregate Litigations

Giovanni Battista Ramello
DiGSPES, University of Eastern Piedmont, Alessandria, Italy
IEL, Torino, Italy

Synonyms

Collective redress; Mass tort litigation; Small claims litigation

Definition

Class action and other forms of aggregate litigation introduce in the legal procedure a powerful means for gathering dispersed interests and channeling them into a type of action in which the different parties concur to promote individual and social interest. They can restore the full working of the legal system, and in addition they can be a powerful device for promoting social welfare when other institutional arrangements seem to be ineffective or inefficient.

Introduction

Class action and other forms of aggregate litigation are the answer to an organizational puzzle in civil procedure dealing with reconciling enforcement of the dispersed victims' rights, the lack of proper incentive for promoting a legal action, and the social interest of producing the public goods of deterrence and, possibly, regulatory change (Ramello 2012). The underlying problem is the twofold partial or total failure of individual litigation and regulation which is essentially explained by the fact that neither institution is able to produce the appropriate incentives for obtaining an appropriate outcome (Dam 1975). In other words, these two production "institutional technologies" are unfitted for the context in which they operate, so that the solution must go by some alternative route.

The economic argument here mirrors that used for explaining the emergence of hierarchies when there is a need to internalize externalities, for example, in the well-known problem in economics of indivisibility in production, which arises in the case of economies of scale (or scope), and makes it impossible to rely on the competitive market for optimal allocation of resources (Edwards and Starr 1984). Indivisibility plays a prominent part in the understanding of industrial organizations and of course likewise affects the market structure. In consequence, the different organizations and multiple forms of enterprises in the market, and of aggregate ventures in the judicial market, should be regarded as institutional solutions designed to achieve adequate productive configurations for specific contexts. The same applies to dispute resolution industry.

It is worth noting that the creation of a hierarchy defines an exclusive right over the specific productive activity. Such a right, in the judicial market, corresponds to a specific legal action and

thus in practice means creating a local monopoly on a particular litigation. This aspect is by no means peripheral to the incentive system in the case of collective redress: it is a prerequisite for being able to assign a property right over the potential rewards of the legal action. Such a right, in its turn, becomes the central element (i.e., the price) for achieving transfer of risk through a contingent fee reward scheme (Eisenberg and Miller 2004, 2013; Sacconi 2011). The party financing the legal action – often the attorney – thus obtains the right to extract a portion of the awarded proceeds as a remuneration for the risk.

Class Versus Aggregate litigation

The currently dominant reference models for aggregate litigation, including class action, are those of the US legal system, whose Rule 20, Rule 23, and Rule 42 of the Federal Rules of Civil Procedure and Section 1407 of Title 28 of the US Code, taken together, introduce various ways of pursuing aggregate litigation in the form of class action, multi-district litigation, formal consolidation, and other solutions, thereby redrawing the boundaries of litigation (ALI 2010).

Rule 23 is the most well known, in that it introduces class action, which has the role of exhausting in a single litigation all possible claims of a predefined population of victims. Among the technicalities of class action, there is also the indirect representation of victims who are unable to join the legal action on their own account (so-called absent parties). The other solutions, in a more fragmentary way, promote collective or coordinated legal actions which, for example, "involve a common question of law or fact [and in which] the court may: (1) join for hearing or trial any or all matters at issue in the actions; (2) consolidate the actions; or (3) issue any other orders to avoid unnecessary cost or delay" (Rule 42a, 2009 edition).

While the specific technical features of each procedural solution are discussed elsewhere (Hensler 2001, 2011; Calabresi and Schwartz 2011), in all cases one of the key criteria for choosing between them is efficiency – meaning the extent to which the aggregation is able to pursue expedition and economy.

Hence, the different forms of aggregation can be compared to the different types of business entities (e.g., public company, joint venture, etc.), whose function is to best exploit the advantages of the hierarchy in different situations. Under this analogy, in the productive organization of the judicial market, class action lies at one extreme, since it exhausts in a single litigation the claims of a broad population of victims who become shareholders in the legal action (essentially a sort of public company). The other solutions occupy intermediate positions, making it possible to exploit some benefits of aggregate litigation even in situations where all the victims cannot join in a single lawsuit, so that a class action is not practicable (and might in fact even be invalidated).

Sketches of History

Although Yeazell (1987) has detected a precursor to class action (and aggregate litigation) in the medieval group litigation of England, class action which is somewhat the benchmark for aggregate litigation was first introduced in the US legal system in 1938, through Rule 23 of the Federal Rules of Civil Procedure. It then took nearly three decades for class action to be fully implemented into the US civil procedure, with the 1966 issuing of the new version of Rule 23 by the Supreme Court. Since then, class action has been fiercely criticized by a number of opponents (Hensler et al. 2000; Klement and Neeman 2004). Despite the negative stances, it has over the years become "one of the most ubiquitous topics in modern civil law" in the USA and nowadays one of "[t]he reason for the omnipresence of class actions lies in [its] versatility" (Epstein 2003, p. 1) which, according to a great many commentators, can make it an effective means for serving justice and efficiency in a broad sense.

The collective litigation system thus continues to operate and to develop, and its utility remains undisputed in the North American judicial system.

The most recent amendment, brought by the Class Action Fairness Act (CAFA 2005; Pub. L. No. 109–2, 119 Stat. 4, 2005), though aimed according to some authors at curbing some of its pernicious features (Lee and Willging 2008), carefully avoided criticizing collective litigation as a whole and in fact reaffirmed its substantive validity, strongly asserting that "class-action law suits are an important and valuable part of the legal system when they permit the fair and efficient resolution of legitimate claims of numerous parties by allowing the claims to be aggregated into a single action against a defendant that has allegedly caused harm" (CAFA 2005, Sect. 2).

In other countries and especially in Europe, aggregate litigations and collective redress systems have recently been introduced. However, local constraints especially derived from the specific legal culture – such as, e.g., the revulsion at accepting the role of the entrepreneurial activity sometime needed in order to trigger the legal action – or simply the political aversions have produced outcomes very distant from the American model, sometime raising substantial concerns about the real effectiveness (Hilgard and Kraayvanger 2007; Baumgartner 2007; Issacharoff and Miller 2012; for a perspective on distinct European countries, see Backhaus et al. (2012) and the contributions therein).

Procedural Features

The first effect of class action and with some variance also of other forms of aggregate litigation is to permit the adjudication of meritorious claims that would otherwise not be litigated due to imperfections in the legal systems (Rodhe 2004). In fact, class action is a legal device employed today for tackling torts in a wide array of cases, including insurance, financial market, and securities fraud in recent times (Pace et al. 2007; Helland and Klick 2007; Porrini and Ramello 2011; Ulen 2011). However, from its inception, class action was infused with a broader political agenda, extending beyond the tort domain to embrace matters such as civil rights (in particular segregation), health protection, consumer protection, environmental questions, and many others (Hensler et al. 2000). This legacy is sometime emerging in other aggregate litigations.

As a whole, collective actions have the effect of altering the balance of power and the distribution of wealth among the various social actors – e.g., firms versus consumers – thereby extending their scope in terms of overall impact on society. All the above elements, taken together, thus play an important role in guiding the legislator's decision of whether (or not) to adopt aggregate litigations, and, ultimately, the battle in favor of or against the introduction of these procedural devices into the different legal systems is played out on a purely political terrain (Porrini and Ramello 2011).

Indeed, it is the procedural technicalities that have for the most part given skeptics grounds for criticizing class action and questioning its ability to be implemented in legal systems different from those where it arose. These are often specious arguments which disregard the simple fact that any "juridical technology" intended to achieve certain outcomes must be adapted, in its design, to the constraints of the target legal system, if it is to provide regulatory solutions that are effective and compatible with its context. The heart of the problem, therefore, consists in opportunely adapting the "legal machinery" to each jurisdictional setting in a manner that obtains the desired results without prejudicing its essential features. These characterizing features can be:

(i) The aggregation of separate but essentially cognate claims, united by design and not by substantive theory
(ii) The indirect representation of absent parties (in the case of class action)
(iii) The provision of entrepreneurial opportunity to an attorney, who thus becomes the main engine of the civil action

Despite different forms of aggregate litigations rely upon distinct features, the common element is that they all try to a great extent to eliminate duplications in related claims, by aggregating in some way the potential claimants into a group. The obvious main consequence is that, by

aggregating in some way similar claims, aggregate litigations increase the possibility to vindicate a tort or however they redress the imbalance which exists between plaintiffs and defendants in several areas of litigation.

The indirect representation, essentially characterizing class action, stems from the fact that the attorney is not appointed directly by each individual claimant, but rather through a specific set of procedures established by law, which essentially rely upon the initiative of a minority among them, and the subsequent acceptance by the judge, to start the trial (Hensler et al. 2000). In fact, the civil action is filed by an individual or a small group of victims assisted by an attorney. The class is then certified by the judge who consequently also "appoints" the attorney as a representative of all the class members (Dam 1975).

It is worth noting that the mere appointment of an attorney does not, of course, per se assure attainment of any efficient outcome, nor does it rule out opportunistic behaviors (Harnay and Marciano 2011). It is only a first step for making the desired outcomes possible and, as usual in tort litigation, demands a well-designed set of incentives for the lawyer in order to work properly (Klement and Neeman 2004; Sacconi 2011).

It is worth reminding that in a sense collective actions bear some similarities – albeit limited to the civil procedure domain – to regulation: in fact, where the judge determines that individual actions may not be sufficiently effective, yet the litigation is in the collective interest, on request of a representation of victims, he or she reallocates the individual rights over that particular prospective litigation. Thus, also in this case, an agent is nominated to represent the interests of a group, but with a narrower scope compared to fully fledged regulation. Here, the indirect representation serves merely to exploit the possibility of aggregating related claims without bearing the costs of searching for and coordinating a huge number – often a "mass"– of potential plaintiffs that would otherwise make bringing the lawsuit unaffordable (Cassone and Ramello 2011).

Finally, there is one last feature that makes collective action possible: it is the creation of a specific entrepreneurial space for the class counsel, who undertakes to identify an unmet demand for justice and, acting self-interestedly, restores access to legal action for the victims. The class counsel is generally driven by the purely utilitarian motives of a "bounty hunter," who offers a service in exchange for recompense (Macey and Miller 1991; Issacharoff and Miller 2012). It is thus a behavior consistent with the paradigm of methodological individualism and which is sometimes regarded with suspicion by those who consider private interests unsuitable for representing the collective interest.

Aggregate litigation thus has the particular merit of aligning the private interest of the case attorney, who seeks to obtain a profit, with that of the victims, who seek redress of the harm and promotion of justice, and with that of society which instead benefits from a system that internalizes the externality. This in fact creates a deterrent to wrongdoing and ultimately may work to minimize the social costs of accidents, in accordance with the Hand's rule (Calabresi 1970). In this light, therefore, the miracle of the invisible hand is again renewed, and the self-interest of the victims and class counsel can play a role of public relevance.

Law and Economics Features

Law and economics is further brought into play when we consider the wider effects of aggregate litigations on the judicial system and on the economic system. In particular, economic science offers two complementary routes for conducting the analysis. The first concerns the manner in which these litigations can serve efficiency and collective welfare; the second provides the analytical framework for representing the legal machinery and studying its workings, thus determining under what conditions and in what way they promote social welfare.

However, for the investigation to be fruitful, we have to specify the initial conditions, i.e., the circumstances under which regulation and individual action are not effective. In other words, we must define the context that gives rise to some shortcomings justifying the introduction of new

legal devices. The conditions may be the following:

- Existence of *fragmented claims*, very often worth less to each plaintiff than the individual litigation cost or which in any case entail a prohibitively costly individual litigation
- Sufficient *homogeneity of claims* for the court to issue a "one size fits all" decision and for the victims to be able to adhere to the collective action
- A *judicial market failure*, as a result of which some claims, no matter how meritorious, are either not brought, so that certain individuals are unable to exercise their rights, or are imperfectly exercised
- A *failure of regulation* which thus does not offer a practicable alternative for resolving the preceding issues

A condition under which aggregate litigation is potentially useful is that where certain rights established by law are not exercised or only imperfectly exercised, due to a misalignment between what is theoretically asserted by the law and the concrete incentives provided to individuals. The solution involves an institutional reorganization to produce a lowering of these costs and/or promote the – sometimes forced – reallocation of the rights.

By thus regarding victims as owners of "property rights" over a specific litigation, whose enforcement may incur costs exceeding the expected individual benefits, we can interpret class action as a system that follows a comparable judicial path to that described for property, aggregating the individuals' rights when their exercise on the judicial market is precluded (or limited) by contingencies which make the net benefit of the action negative (Ramello 2012).

In general, these contingencies arise from the aforesaid fragmentation and its attendant coordination costs, from the limited size of the individual damages (so-called small claims), and also from the existence of asymmetries between the would-be plaintiffs and defendant (i.e., availability of information, capacity to manage the litigation risk, access to financial resources, and more).

Creating a pool of rights thus enables victims to access a less costly litigation technology and thereby pursue justice. The productive efficiency of a static character concerns the overall production of "justice," on the demand and supply sides, since on the judicial market, both jointly concur to its production, albeit for different reasons. Aggregate litigations in fact allow a so-called judicial economy to emerge, which on the demand side, e.g., through aggregation of small claims, produces economies of scale in litigations that cause individual costs to decrease with increasing number of plaintiffs (Bernstein 1977). On the supply side, there is likewise a reduction in costs if the aggregation permits overall savings in resources compared to multiple individual actions, provided though that the savings afforded by aggregation are not offset by an increase in the number of lawsuits.

There is, then, a third level of efficiency connected with the economic nature of aggregate litigation and which has the purpose of aligning different interests to achieve the previously stated goal. In effect, the system, if properly applied, has to introduce a set of distinct incentives which together concur to produce three different outputs: a profit for the attorney, redress of the harm for the victims, and deterrence of wrongdoing (thereby minimizing the social cost) for society.

Using the traditional categories of economic analysis and with special reference to class action, Cassone and Ramello (2011) disentangle the "productive" roles of the various actors taking part in the litigation. In other words, the role of these legal procedural devices reconciles the conjoined individual interests of victims with the collective interest of society, by passing through the private interest of the attorneys. It thus has the nature of a *private good* for the attorney, who takes on the entrepreneurial role of setting in motion or managing the collective action, which is in its turn aimed at obtaining redress of the harm (Dam 1975). Though this ultimately has an effect on each victim, it can only be produced as a local public good for the cohort of all victims and thus takes the form of a *club good* (Cassone and Ramello 2011). Finally, the transfer of the cost of the wrongdoing from the victims to the injurer

has the consequence of reestablishing a higher level of deterrence, thereby resulting in production of a *public good* (Eisenberg and Engel 2014). This deterrence, it is worth noting, pertains to what is generally termed dynamic efficiency, since its production in a given time frame is also instrumental to the intertemporal optimal production of other goods. In fact, besides the usual production of deterrence, as in general done by tort law, there can also be the production of inputs for regulation, thus establishing a causal relation between litigation and regulatory rule-making. In this respect, aggregate litigations have the additional feature of producing information externalities and consensus among a broad panel of individuals acting as a proxy for the society that serve as direct inputs to regulation. Then the litigation becomes a sort of R&D laboratory, in which plaintiffs act as a proxy for society and the judicial solution serves as a prototype for regulatory change (Arlen 2010; Ramello 2012).

Cross-References

► Mass Tort Litigation: Asbestos

References

ALI (2010) Principle of the law of aggregate litigation. American Law Institute, Philadelphia

Arlen J (2010) Contracting over liability: medical malpractice and the cost of choice. Univ Penn Law Rev 158:957–1023

Backhaus JG, Cassone A, Ramello GB (2012) The law and economics of class actions in Europe. Lessons from America. Edward Elgar, Cheltenham

Baumgartner SP (2007) Class actions and group litigation in Switzerland. Northwest J Intl Law Bus 27(2):301–349

Bernstein R (1977) Judicial economy and class action. J Legal Stud 7(2):349–370

CAFA 2005 is an acronm for Class Action Fairness Actas written in the text the first time. Hence it does not need reference

Calabresi G (1970) The cost of accident. Yale University Press, New Haven

Calabresi G, Schwartz KS (2011) The costs of class actions: allocation and collective redress in the US experience. Eur J Law Econom 32:169–183

Cassone A, Ramello GB (2011) The simple economics of class action: private provision of club and public goods. Eur J Law Econ 32:205–224

Dam KW (1975) Class actions: efficiency, compensation, and conflict of interest. J Legal Stud 4:47–73

Edwards BK, Starr RM (1984) A note on indivisibilities, specialization, and economies of scale. Am Econom Rev 77:192–194

Eisenberg T, Engel C (2014) Assuring civil damages adequately deter: a public good experiment. J Empir Legal Stud 11:301–349

Eisenberg T, Miller GP (2004) Attorney fees in class action settlements: an empirical study. J Empir Legal Stud 1:27–78

Eisenberg T, Miller G (2013) The english vs. the American rule on attorneys fees: an empirical study of attorney fee clauses in publicly-held companies' contracts'. Cornell Law Rev 98:327–382

Epstein RA (2003) Class action: aggregation, amplification and distorsion. University of Chicago, Legal Forum, pp 475–518

Harnay S, Marciano A (2011) Seeking rents through class actions and legislative lobbying: a comparison. Eur J Law Econ 32:293–304

Hensler DR, Dombey-Moore B, Giddens E, Gross J, Moller E, Pace M (eds) (2000) Class action dilemmas. Pursuing public goals for private gain. Rand Publishing, Santa Monica/Arlington

Helland E, Klick J (2007) The trade-off between regulation and litigation: evidence from insurance class actions. J Tort Law 1:1–24

Hensler DR (2001) Revisiting the monster: new myths and realities of class action and other large scale litigation. Duke J Comp Int Law 11:179–213

Hensler DR (2011) The future of mass litigation: global class actions and third-party litigation funding. George Wash Law Rev 79:306–323

Hilgard MC, Kraayvanger J (2007) Class action and mass action in Germany. International Bar Association Litigation Committee Newsletter, September, pp 40–41

Issacharoff S, Miller GP (2012) Will aggregate litigation come to Europe? In: Backhaus JG, Cassone A, Ramello GB (eds) The law and economics of class actions in Europe. Lessons from America. Edward Elgar, Cheltenham

Klement A, Neeman Z (2004) Incentive structures for class action lawyers. J Law Econ Org 20:102–124

Lee EG, Willging TE (2008) The impact of the class action fairness act on the federal courts: an empirical analysis of filings and removals. Univ Penn Law Rev 156:1723–1764

Macey JR, Miller GP (1991) The plaintiffs' Attorney's role in class action and derivative litigation: economic analysis and recommendations for reform. Univ Chicago Law Rev 58:1–118

Pace NM, Carroll SJ, Vogelsang I, Zakaras L (2007) Insurance class actions in the United States. RAND, Sanata Monica

Porrini D, Ramello GB (2011) Class action and financial markets: insights from law and economics. J Fin Econ Policy 3:140–160

Ramello GB (2012) Aggregate litigation and regulatory innovation: another view of judicial efficiency. Int Rev Law Econ 32(1):63–71

Rodhe DL (2004) Access to justice. Oxford University Press, Oxford/New York
Sacconi L (2011) The case against lawyers' contingent fees and the misapplication of principal-agent models. Eur J Law Econ 32:263–292
Ulen TS (2011) An introduction to the law and economics of class action litigation. Eur J Law Econ 32:185–203
Yeazell SC (1987) From medieval group litigation to the modern class action. Yale University Press, New Haven

Climate Change Remedies

Donatella Porrini
Department of Management, Economics, Mathematics and Statistics, University of Salento, Lecce, Italy

Abstract

The law and economics analysis of the climate change remedies has been focused on the question of which would be the policy instrument most suited to provide incentives to reduce greenhouse gas emissions. The literature focuses mainly on the comparison of carbon taxes and emission trading scheme. But a relevant role can be played by financial and insurance instruments, especially considering the adaptation and mitigation strategies. Finally, another instrument is considered, largely used to internalize other environmental externalities but still not so much analysed for climate change, the liability system.

Definition

Over the last centuries, climate change has become a very important issue all over the world. The change in climate corresponds to an increase in the earth's average atmospheric temperature, which is usually referred to as global warming.

In response to scientific evidence that human activities are contributing significantly to global climate change, and particularly the emissions of greenhouse gas (GHG) emissions, decision-makers are devoting considerable attention to find remedies to reduce the consequences in terms of climate change.

The Concept of "Economic Global Public Goods"

Dealing with climate change implies the concept of "economic global public goods" that can be defined as goods with economic benefits that extend to all countries, people, and generations (Kaul et al. 2003).

First of all, the emissions of GHG have effects on global warming independently of their location, and local climatic changes are completely linked with the world climate system.

In addition, the effects of GHG concentration in the atmosphere on climate are intergenerational and persistent across time.

The fact that climate change is clearly "global" in both causes and consequences implies that, on one side, we cannot determine with certainty both the dimension and the timing of climate change and the costs of the abatement of emissions, on the other side, it emerges a relevant equity issue among countries because industrialized countries have produced the majority of GHG emissions, but the effects of global warming will be much more severe on developing countries.

About this last point, the countries that have more responsibilities will face less consequence in the future and vice versa. So it is a global issue to decide the distribution of emission reductions among countries and how the costs should be allocated, taking into account the differences among countries characterized by high- or low-income, high- or low- emissions level, and high and low vulnerability.

Climate change is going to generate natural disasters, meaning events caused by natural forces that become "man-made" disasters, meaning events associated with human activities, given the role of greenhouse gases emitters. More precisely, we can speak of "unintended man-made" disasters originated by global warming (Posner 2004, p. 43).

The rising costs associated with climate change effects pose serious challenges to governments to adopt efficient strategies to manage the increasing

economic consequences, and governments are facing the issue to introduce policies to tackle the causes and combat all the effects of greenhouse gas emissions.

Dealing with global public goods, the choice of environmental policies requires a global coordination (Nordhaus 2007). But, in any case, it is difficult to determine and reach agreement on efficient policies because economic public goods involve estimating and balancing costs and benefits where neither is easy to measure and both involve major distributional concerns. As a consequence, it is necessary to reach through governments to the multitude of firms and consumers who make the vast number of decisions that affect the ultimate outcomes.

Carbon Tax and Emission Trading Scheme

The policy instruments that are mainly implemented as remedies against climate change are carbon tax and emission trading scheme (ETS).

A carbon tax is a particular levy on GHG emissions generated by burning fuels and biofuels, such as coal, oil, and natural gas. It is generally introduced with the main goal to level the gap between carbon-intensive (i.e., firms based on fossil fuels) and low carbon-intensive (i.e., firms that adopt renewable energies) sectors.

A carbon tax provides a strong incentive for individuals and firms to adjust their conduct, resulting in a reduction of the emissions themselves because the relative price of goods and services changes. Hence, by decreasing fuel emissions and adopting new technologies, both consumers and businesses can reduce the entire amount they pay in carbon tax.

An emission trading scheme (ETS) is an instrument based on an agreement that sets quantitative limits of emissions and the allocation of emission, allowing the trade in order to minimize abatement costs. At the beginning the allocation of permits can be set through either an auction or a grandfather allocation. Under an auction, government sells the emission permits, whereas under the grandfather rule, the allocation of emission permits is based on historical records.

An ETS is defined as a quantity-based environmental policy instrument. It is also called cap and trade because it is characterized by the allowable total amount of emissions (cap) and the right to emit that becomes a tradable commodity. Under an ETS system, prices are allowed to fluctuate according to market forces.

On the other hand, carbon taxes are defined as price-based policy instruments for the correlated effects to increase the price of certain goods and services, thereby decreasing the quantity demanded.

An emission trading system may efficiently give the incentive to decrease the emissions wherever abatement costs are lowest with positive effects beyond the national borders. As costs associated with climate change have no correlation with the origin of carbon emissions, the rationale for this policy approach is that an emission trading system allows to fix a certain environmental outcome and the companies are called to pay a market price for the rights to pollute regardless of where there will be the benefits. This is the reason why an emission trading system is suitable for international environmental agreements, such as the Kyoto Protocol, and also for the characteristic that a defined emission reduction level can be easily agreed between states.

Emission trading can be an advantage for private industry because, by decreasing emissions, firms can actually profit by selling their excess GHG allowances, in a way that such a market of permits could potentially drive emission reductions below targets.

A system of ETS entails significant transaction costs, which include search costs, such as fees paid to brokers or exchange institutions to find trading partners, negotiating costs, approval costs, and insurance costs. Conversely, taxes involve little transaction cost over all stages of their lifetime.

Carbon taxes are economic instruments that works dynamically offering a continuum incentive to reduce emissions. In fact, technological and procedural improvements and their subsequent efficient diffusion lead to decreasing tax payment. On the other hand, trading systems will adjust

when the emission goals are easier to meet, so that in this case a decreasing demand of permits causes a reduction in their price but not as rapidly as taxes.

The law and economics literature describes as alternative instruments carbon tax and tradable permit system, the former as a price control instrument and the latter as a quantity control one.

Many contributions compare the relative performance of price and quantity instruments under uncertainty, starting with the seminal contribution of Weitzman (1974). For example, Kaplow and Shavell (2002) deal with the standard context of a single firm producing externality; moreover, they consider the case of multiple firms that jointly create an externality, concluding with the superiority of taxes to permits.

In the case of climate change, there are arguments for price controls. The first point is that climate change consequences are uncertain because it is not the level of annual emissions that matters, but rather the total amount of GHG that have accumulated in the atmosphere. The second point is that "while scientists continue to argue over a wide range of climate change consequences, few advocate an immediate halt to further emission" (Pizer 1999, p. 7).

Even if a carbon tax is preferable to an ETS scheme in terms of social costs and benefits, this policy obviously faces political opposition. Private industry opposes carbon taxes because of the transfer of revenue to the government; environmental groups oppose carbon taxes for an entirely different reason: they are unsatisfied with the prospect that a carbon tax, unlike a permit system, fails to guarantee a particular emission level.

The Role of Financial and Insurance Instruments

To face climate change economic consequences, a role can be assigned also to private sector to stimulate the reduction of the probability of catastrophic losses and to manage economically large-scale disaster risks. In this sense a relevant part can be played by the financial and insurance products that are based on mechanisms to manage the economic consequences of risk, including the threat posed by natural hazards.

With the typical insurance contract, for example, individuals and companies protect themselves against an uncertain loss by paying an annual premium toward the pool's expected losses. The insurer holds premiums in a fund that, along with investment income and supplementary capital (where necessary), compensates those that experience losses.

First of all, climate change consequences are insured through the coverage of the risks that insurance companies accept from their customers, since policies already include the provision of the economic consequences of changes in the intensity and distribution of extreme weather events and of the resulting risk of catastrophic property claims (Porrini 2011).

The supply of this kind of products, that are the core business of the insurance industry, experiences some problems.

First, climate change's relationship to global weather patterns increases the potential for losses so large that they threaten the solvency of the insurance companies.

Second, uncertainties in assessing climate change's impacts are high, affecting property and casualty, business interruption, health, and liability insurance, among others. As a result, insurers could charge a significantly higher premium or, in certain cases, avoid to supply this kind of policies.

Third, many climate change-related risks may be correlated, creating a skewed risk pool and exacerbating the risk of extremely large losses, and that some of these risks are not well distributed across existing insureds.

Beyond the problems of insurability, financial and insurance market provide for other kind of products. Examples are "compensation funds," such as special government disaster funds, to promote framework of contingency measures to tackle climate change consequences. These funds, created in connection with a regulatory system mainly to cover environmental damage and victims' compensation, can be financed by a taxation system or by a firm's contribution system. The main example is the Superfund in the

United States, connected with the regulatory system by Environmental Protection Agency (EPA).

Other examples are products characterized by ex ante commitment of financial resources, such as the so-called "financial responsibility" instruments. This term defines all the tools that require polluters to demonstrate ex ante sufficient financial resources to correct and compensate for environmental damage that may arise through their activities.

In its common application, financial responsibility implies that the operation of hazardous plants and other business is authorized only if companies can prove that future claims will be financially covered, for example, through letters of credit and surety bonds, cash accounts and certificates of deposit, self-insurance, and corporate guarantees.

Generally, financial responsibility may be complementary, sometimes mandatory, to the legislation on environmental accidents. In its different applications, it has a common motivation: to ensure the future internalization of the costs in order to indemnify the victims and discourage different forms of environmental deterioration.

On a law and economics point of view, financial responsibility can be defined as (potentially) efficient instruments to correct the asymmetric information issue. First of all, there is an incentive for the financial institutions to check that the companies are taking adequate preventive measures. Secondly, the companies are motivated to take precautions because financial responsibility guarantees that the expected costs of environmental risks appear on their balance sheet and business calculation (Feess and Hege 2000).

There are also alternative risk transfer products. A first kind of products is catastrophe bonds, consisting in securitizing some of the risk in bonds, which could be sold to high-yield investors. The cat bonds transfer risk to investors that receive coupons that are normally a reference rate plus an appropriate risk premium. By these products, financial institutions may limit risk exposure transferring natural catastrophe risk into the capital markets.

Weather derivatives are another kind of financial instrument used to hedge against the risk of weather-related losses. Weather derivatives pay out on a specific trigger, e.g., temperature over a determined period rather than proof of loss. The investor providing a weather derivative charges the buyer a premium for access to capital, but if nothing happens, then the investor makes a profit.

With all this kind of insurance and financial products, it is possible to reach some efficiency goals. First of all, they give the possibility to stimulate ex ante preventive measure and to economically compensate ex post the victims. The second goal is the availability of extra capital for recovery that comes from financial markets. Finally, the accuracy and the resolution of hazard data and the likely impacts on climate change may improve with the involvement of financial market forecast ability.

The Mitigation and Adaptation Strategies

The challenge of reducing in the future the consequences of climate change is often framed in terms of two potential strategies: adaptation and mitigation. Mitigation involves lessening the magnitude of climate change itself; adaptation, by contrast, involves efforts to limit the vulnerability to climate change impacts through various measures, while not necessarily dealing with the underlying cause of those impacts.

"Mitigation" indicates any action taken to permanently eliminate or reduce the long-term risk and hazards of climate change to human life. A definition can be "An anthropogenic intervention to reduce the sources or enhance the sinks of greenhouse gases" (IPCC 2001).

"Adaptation" refers to the ability of a system to adjust to climate change to moderate potential damage, to take advantage of opportunities, or to cope with the consequences. A definition can be "Adjustment in natural or human systems to a new or changing environment" (IPCC 2001).

Mitchell and Tanner (2006) defined adaptation as an understanding of how individuals, groups, and natural systems can prepare for and respond to changes in climate. According to them, it is crucial to reduce the vulnerability to climate change.

While mitigation tackles the causes of climate change, adaptation tackles the effects of the phenomenon. The potential to adjust in order to minimize negative impact and maximize any benefits from changes in climate is known as adaptive capacity. A successful adaptation can reduce vulnerability by building on and strengthening existing coping strategies.

In general, the more mitigation there is, the less will be the impacts to which we will have to adjust and the less the risks for which we will have to prepare. Conversely, the greater the degree of preparatory adaptation, the less may be the impacts associated with any given degree of climate change.

The idea is that less mitigation means greater climate change effects, and consequently more adaptation is the basis for the urgency surrounding reductions in greenhouse gases. The two strategies are implemented on the same local or regional scale and may be motivated by local and regional priorities and interests, as well as global concerns. Mitigation has global benefits, although effective mitigation needs to involve a sufficient number of major GHG emitters to foreclose leakage. Adaptation typically works on the scale of an impacted system, which is regional at best, but mostly local, although some adaptation might result in spillovers across national boundaries.

Climate mitigation and adaptation should not be seen as alternatives to each other, as they are not discrete activities but rather a combined set of actions in an overall strategy to reduce GHG emissions. The challenge is to define an efficient mix of government policy interventions to provide the right incentives to invest in cost-effective preventive measures to reduce the final cost of disasters. The target is to tackle the consequences of climate change by mitigation, through the promotion of ways to reduce greenhouse gas emissions and make society to adapt to the impacts of climate change, by promoting the effective limitation and management of risks from extreme weather-related hazards.

On a law and economics perspective, generally, private contracting has been recognized as a significant and potentially effective means of influencing private actors' behavior and even as a form of environmental policy instrument. So the financial and insurance products, that we have above analyzed, have significant potential to influence the behavior of individuals through its contracting contents, and this implies that the financial markets can play a role within the mitigation and adaptation policies.

For example, insurance companies may offer differential premiums to customers depending on the customers' level of protection from loss caused by weather-related disasters with an opportunity for insurers to reduce their own overall and maximum possible loss exposure while promoting communities' overall resilience in the face of climate change's impacts. Moreover, financial products can include arrangements intended to bring needed capital that will reduce the risk posed by future climate-related hazards to those who are most likely to be in peril.

Financial and insurance products could affect incentives for individuals to address climate change seeking mechanism to facilitate mitigation of GHG and adaptation to the inevitable impacts of climate change. Additionally, financial institutions are motivated to take significant actions aimed at mitigating overall societal greenhouse gas emissions and increasing adaptive capacity because these actions would reduce overall uncertainty and decrease people and business' potential exposure to catastrophic risks in excess of their capacity.

Conclusive Remarks on a Future Climate Change Liability System

The law and economics analysis of the climate change remedies has been focused on the question of which would be the policy instrument most suited to provide incentives to industry and other sources to reduce greenhouse gas emissions. And the literature is still giving attention mainly to the comparison of carbon taxes and emission trading scheme (Nordhaus 2006).

Not so much attention has been addressed to another instrument to provide incentives to polluters to reduce emissions, largely used to internalize other environmental externalities, the liability system. With "liability" we intend the possibility of applying national tort law to the damage caused by climate change and the possibility for holding states liable under international law if emissions originating from a country were to cause damage to the citizens of other nations.

Even if it seems that the application of a liability system to climate change is merely a theoretical issue, in reality more and more public authorities or individuals have tried to sue large emitters of GHG, and, in some cases, claims were directed against governmental authorities for failure to take measures to reduce emissions of greenhouse gases.

As an example, in 2002, the small island state Tuvalu threatened to take the United States and Australia to the International Court of Justice as a result of their failure to stabilize GHG emissions, thus causing the melting of ice caps which consequently leads to a rise in sea levels which threatened its territory. Although for a change in government the application was never made, this example demonstrates the way in which international law could be used to impose liability for climate change-related harm.

Beyond this specific case, most of these claims would probably not qualify as liability suits in the strict sense, since it is usually not compensation for damage suffered that is asked by the plaintiffs, but rather injunctive relief in order to obtain a reduction of greenhouse gasses. Moreover, most of the claims brought so far, mainly in the United States, were either not successful, were withdrawn, or have not yet led to a specific result.

On a law and economics point of view liability is not only an instrument "to obtain compensation for damages resulting from climate change (the more traditional liability setting) but equally are looking at the question to what extent civil liability and the courts in general may be useful to force potential polluters (or governmental authorities) to take measures to reduce (the effects of) climate change" (Faure and Peeters 2011, p. 10).

A liability system could also play a role in mitigating climate change, and a question is open to what extent it is useful to use the civil liability system to strive for a mitigation of greenhouse gas emissions in addition to the existing framework which largely relies on carbon tax and emission trading systems.

References

Faure M, Peeters M (2011) Climate change liability. Edward Elgar, Cheltenham

Feess E, Hege U (2000) Environmental harm, and financial responsibility. Geneva Pap Risk Insur Issue Pract 25:220–234

IPCC (2001) Climate change 2001: synthesis report. In: RT Watson and the Core Team (eds) A contribution of working groups I, II, III to the third assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge/New York

Kaplow L, Shavell S (2002) On the superiority of corrective taxes to quantity, American Law and Economics Review, 4, pp. 1–17

Kaul I, Conceicao P, Le Goulven K, Mendoza RU (2003) How to improve the provision of global public goods. In: UNDP (ed) Providing global goods – managing globalization. Oxford University Press, New York

Mitchell T, Tanner TM (2006) Adapting to climate change: challenges and opportunities for the development community. Tearfund, Middlesex

Nordhaus DW (2006) After Kyoto: alternative mechanisms to control global warming. FPIF discussion paper

Nordhaus DW (2007) To tax or not to tax: alternative approach to slowing global warming. Rev Environ Econ Policy 1:26–40

Pizer W (1999) Choosing price or quantity controls for greenhouse gases, resources for the future. Climate Issue Brief no 17

Porrini D (2011) The (potential) role of insurance sector in climate change economic policies. Environ Econ 2(1):15–24

Posner R (2004) Catastrophe. Oxford University Press, New York

Weitzman ML (1974) Prices vs. quantities. Rev Econ Stud 41(4):477–491

Clinical Trials

► Human Experimentation

Coase, Ronald

Herbert Hovenkamp
The College of Law, The University of Iowa, Iowa City, IA, USA

Abstract

Ronald Coase (1910-2013) was a British born and trained economist who moved to the United States in 1951. He spent most of his career at the University of Chicago. Coase's principal contributions addressed the fact that moving resources through the economy by means of transactions is costly – an idea that he introduced in The Nature of the Firm (1937) and developed further in The Problem of Social Cost (1960). Over his career Coase argued in numerous papers that if transaction costs are modest, private bargaining is often better than legislation or taxation as devices for settling resource conflicts. His work was highly influential in the development of movements away from regulation and back to more market-centric devices for managing the private economy. Coase won the Nobel Prize in economics in 1991.

Biography, Impact and Legacy

Ronald Harry Coase (1910–2013) was born in suburban London. His father worked for the British post office as a telegraph operator, as did his mother until marriage. Coase attended the University of London and then the London School of Economics, receiving a Bachelor of Commerce degree in 1932. After lectureships at Dundee and Liverpool, he returned to LSE from 1935 to 1951. He then moved to the United States to the State University of New York at Buffalo. In 1958 he went to the University of Virginia, and in 1964 to the University of Chicago. For a time he was editor of the *Journal of Law and Economics*. Coase received the Nobel Prize in Economics in 1991. By his own admission, he did not like mathematics – a fact that set him apart from most of the economists of his generation.

Coase was hardly the most prodigious writer among Nobel laureate economists, but what he wrote was highly influential. Indeed, in a real sense he may be called the father of the discipline of law and economics. His reputation rests heavily on two articles. "The Nature of the Firm" (Coase 1937) was written during the period 1932–1934 while Coase was an assistant lecturer at the School of Economics and Commerce, Dundee, Scotland (Coase 1994). He wrote "The Problem of Social Cost" (Coase 1960) while he was at the University of Virginia.

Coase has stated that "The Nature of the Firm" was conceived in 1931 and essentially finished in 1934. He was in his early twenties and just beginning his academic career. His article was intended to address a different issue than the ones that eventually made it prominent. Marginalists since the great industrial economist Alfred Marshall at Cambridge were troubled about why a single industry contains firms of different sizes and structures. For example, if fixed costs were at all substantial, one might expect the market to be taken over by a single firm, which would have lower per unit costs than any rival. In the highly influential eighth edition of his *Principles of Economics* (1920), Marshall developed the idea of the "representative firm," a hypothetical mature firm with "normal" cost characteristics, although individual firms in various stages of development could vary. Marshall never specified precisely what made a firm "representative," and identifying it was like identifying the "representative" tree in a forest (Marshall 1890; 1920).

This analytically unsatisfactory theory led to criticism and attempts at refinement. One influential critique was American economist John Maurice Clark's book on fixed costs (Clark 1923), which addressed the problem in terms of diverse technologies and pricing strategies: scale economies do not produce a single firm because firms are not identical. The subsequent rise of theories of product differentiation made the idea of a "representative" firm obsolete in microeconomics,

Ben V. & Dorothy Willie Professor of Law and History, University of Iowa. Thanks to Erik Hovenkamp and Robert T. Miller for commenting on a draft

although it retained more traction in macroeconomics, particularly in Keynes. Firms in differentiated markets can have quite different sizes and structures. They compete by appealing to divergent consumer tastes.

In 1928 Lionel Robbins, head of the London School of Economics, strongly criticized Marshall's concept of the representative firm for failing to apply the very marginalist rigor that Marshall himself advocated (Robbins 1928). Arthur Cecil Pigou, Marshall's successor at Cambridge, developed the idea of the "equilibrium firm," arguing that a firm will expand when its marginal cost is lower than the market's supply price but contract when it is higher (Pigou 1928). The equilibrium firm is one whose marginal cost just equals the market supply price. In 1931 Cambridge economist E.A.G. Robinson added in *The Structure of Competitive Industry* that "management costs" must also be considered in any question about firm size and structure (Robinson 1931; Hovenkamp 2011). So Coase was not writing on a clean slate.

This history explains Coase's strange paean to Marshall in the opening paragraphs of "The Nature of the Firm." Coase stated his intent to use "two of the most powerful instruments of economic analysis developed by Marshall, the idea of the margin and that of substitution, together giving the idea of substitution at the margin" (Coase 1937). By 1937 the ideas of marginalism and substitution to equilibrium had become conventional in economics. They were not worth mentioning, except that Coase was pointing out a gap in Marshall's approach. Coase then observed that marginal cost includes all relevant incremental costs, including what he termed "marketing costs," by which he meant "the costs of using the price mechanism." The term "transaction costs," for which Coase is now popularly associated, did not appear in this article.

Coase's highly elegant model argued that for every production or distribution decision, a firm compares alternative approaches, including purchase on the market as an alternative to internal production, by various means. Internal production, internal management, and use of external markets are all costly. The firm's management selects the alternative that maximizes firm value. The aggregate of these decisions accounts fully for the firm's size and "shape" – that is, the variety of markets in which it operates and the extent of its vertical integration. The elegance of Coase's argument lay not only in its simplicity but also its enormous range, extending far beyond vertical integration itself to such questions as whether to differentiate one's product or use more or less centralized governance, equity or debt financing, and the like. In the process "The Nature of the Firm" developed a powerful theme that came to dominate Coase's work – namely, a property right and a contract are simply alternative ways of getting something done. For example, an automobile maker's decision whether to build its own spark plugs or purchase them is simply a choice between property and contract.

Over his career Coase repeatedly criticized "blackboard" economists who abstracted from reality, repeatedly calling for more empirical research (e.g., Coase 1992). But the empirical research that went into "The Nature of the Firm" is minimal. Coase visited a few firms, conducted a very few interviews, and overheard some phone calls about procurement. His theory was purely analytic in the British tradition, assuming how a rational actor would select among alternative production decisions.

"The Nature of the Firm" lay ignored for 30 years after its publication, with leading texts on industrial organization not even mentioning it (e.g., Bain 1959). In 1942 the prominent economist and public intellectual Kenneth E. Boulding wrote an article discussing the leading literature on the theory of the firm over the preceding 10 years, but did not cite Coase's article (Boulding 1942). It was finally rediscovered after "The Problem of Social Cost" was published in 1960 (Cheung 1983).

In 1959 Coase published an article arguing that an auction-style market would be a better way to allocate radio spectrum than the largely political arrangements currently in use: ". . .it is not clear why we should have to rely on the Federal Communications Commission rather than the ordinary pricing mechanism to decide whether a particular frequency should be used by the police, or for

a radiotelephone, or for a taxi service. . .” (Coase 1959). That article contained this insight that came out of “The Nature of the Firm” but became the basis of “The Problem of Social Cost” a year later. Speaking of a cave, Coase noted that the law of property determines who owns it. However,

> . . .the law merely determines the person with whom it is necessary to make a contract to obtain the use of the cave. Whether the cave is used for storing bank records, as a natural gas reservoir, or for growing mushrooms depends, not on the law of property, but on whether the bank, the natural gas corporation, or the mushroom concern will pay the most in order to be able to use the cave. (Coase 1959, at 25)

The principal purpose of governmental spectrum allocation, Coase observed, was to prevent interference that occurred when spectrum assignments conflicted with one another. Coase introduced the case of Sturges v. Bridgeman (1879), a nuisance dispute involving a physician who shared a building with a confectioner. The thumping of the confectioner’s mechanical mortar and pestle interfered with the physician’s use of his stethoscope. Coase pointed out that neither *Sturges* nor the spectrum case involved conflicts between a wrongdoer and a victim. They merely represented inconsistent property interests. In a well-functioning market, the interest would go to the person who was willing to pay the most for it.

Coase elaborated on this theme a year later in “The Problem of Social Cost” (Coase 1960). His foil was no longer the FCC but rather Pigou, who had died the previous year. Pigou was the first neoclassical economist to write extensively about how the costs of moving resources should be factored into economic analysis, although his concept of “costs of movement” was more inclusive than Coase’s “transaction costs.” Pigou argued that in cases involving multiple, unorganized users of rivalrous resources, individuals would tend toward excessive use. In such cases the state should intervene with taxes or regulations designed to encourage efficient use. One example that Pigou gave and Coase discussed was the factory that belched smoke, injuring downwind landowners. Clean air was the resource in question. To the extent the factory did not bear the full social cost of dirty air it would overpollute. Pigou argued that the factory should be given legal liability so as to reduce or eliminate the smoke, or else assessed a tax that was “equivalent in money terms to the damages it would cost” (Pigou 1932, Chapter 9). Coase argued that it was incorrect to think of the factory as the “wrongdoer” and the property owners as victims. Both performed useful social activities that were simply inconsistent uses of land. His second point was that without transaction costs, private bargaining would address the problem, not necessarily by shutting the factory down, but rather by assigning the right to whoever valued it most highly.

Coase did not invent the term “Coase Theorem.” That credit belongs to George J. Stigler, Coase’s colleague at Chicago. Stigler also provided this definition: “Under Perfect Competition Private and Social Costs will be Equal” (Stigler 1966, p. 113). The definition was probably intended to capture Coase’s differences with Pigou.

Stigler’s initial definition caught Coase’s insights very poorly. Coase’s paper had virtually nothing to do with perfect competition. The markets in “The Problem of Social Cost” are largely bilateral monopolies, and Coase readily acknowledged that the price at which legal entitlements in such markets are transferred is indeterminate. Under perfect competition prices are at marginal cost. Finally Stigler’s definition trivializes the Coase Theorem by turning it into a minor and fairly obvious corollary of the First Welfare Theorem, which was already well known when “The Problem of Social Cost” was published (Blaug 2007). Coase corrected Stigler’s statement to say “with zero transaction costs private and social costs will be equal” (Coase 1988b, p. 158). Stigler later revised his definition to state “when there are no transaction costs the assignments of legal rights have no effect upon the allocation of resources among economic enterprises. . .” (Stigler 2003, at 77).

As formalized, the Coase Theorem is said to have two parts, or perhaps two different applications, which are not mutually exclusive. First, an “efficiency” thesis states that if transaction costs are zero, then the initial allocation of a right is irrelevant to efficiency because the right will be

traded to its highest value user. The final allocation maximizes private value among the bargainers. It also maximizes social value, *provided* that no outsider to the bargain is adversely affected – that is, there are no negative externalities. Second, an "invariance" thesis states that if transaction costs are zero, then where the right ends up is invariant to the underlying legal rule that creates it – "irrespective of the initial assignment of rights" (Coase 1992). One limitation is that the right must be "alienable," meaning that the parties can contract around it through settlement. For example, in *Sturges* it does not matter whether Bridgman's mortar and pestle is or is not declared a nuisance. Whether the machine is shut down depends entirely on whether Sturges values the right to be free of the noise more than Bridgeman values the right to use the machine. One problem with "inalienable" legislated rights is that the parties cannot bargain around them. For example, if a zoning law prohibited the use of Bridgeman's machine, then the parties could not bargain to the efficient solution if use of the machine was more valuable than the interference it caused. At least for biological actors, the endowment effect can undermine the invariance thesis if an actor's willingness to accept for a particular right is greater than his willingness to pay (Hovenkamp 1990; Kahneman et al. 1990).

Writing about the Coase Theorem has been voluminous, making "The Problem of Social Cost" the most cited law review article of all time. It drew an almost immediate response in tort and property law, two areas where the infant law and economics movement cut its teeth. In 1964 University of Chicago law professors Walter Blum and Harry Kalven acknowledged its importance in an article on tort liability. They observed, however, that the actors in Coase's account were neighbors well aware of the accident possibilities before they occurred. The theorem would not work for automobile accidents, however, because prior to the accident the parties would not be in a position to negotiate over such issues as right of way (Blum and Kalven 1964; see Medema 2013).

By contrast, Guido Calabresi developed an alternative that relied on objective criteria for determining who would have won a bargain had the parties been able to negotiate. In such cases liability should be assigned to the "least cost avoider" (Calabresi 1970). From there the debate spread into numerous areas, including questions such as when strict liability was a more efficient tort rule than negligence. In property law the literature considered whether common law rules such as nuisance or else private restrictive covenants were effective alternatives to zoning (e.g., Ellickson 1973).

Another issue was the choice between "alienable" rules that could be privately bargained and "inalienable" rules that could not be (Calabresi and Melamed 1972). Generally speaking, private injunction rules with alienable entitlements set up a mechanism like the one Coase contemplated. The rule creates a property interest in a plaintiff, if entitled to the injunction, but permits the parties to bargain around it. By contrast, a pure damages rule permits the conduct to continue but may require one person to pay the other an amount determined by the court. Once again, however, the parties are free to negotiate their own private arrangement. An "inalienability" rule, by contrast, assigns the right in one way and prohibits the parties from changing it by private agreement.

Generally speaking, inalienability rules make the most sense when transaction costs are high, meaning that the parties are unlikely to reach the efficient bargain, *provided* the state can by objective means determine which party would have won the right in a free bargain. For example, the common law rule requiring cars to yield to trains at grade crossings ordinarily creates an inalienability rule. When both are speeding toward the intersection, the car and the train are not in a good position to bargain over the right of way. Further, the costs of stopping and restarting are much higher for the train than for the car. So the state assigns the right of way to the train.

The choice between injunction rules and damages rules is more problematic. One view is that injunction rules are preferred when transaction costs are low and the parties are likely to bargain to an efficient result. By contrast, damages rules are superior when value determination is complex, perhaps because multiple parties are involved or there might be holdouts. One critique

of this view is that it implicitly assumes that the court is a better decision maker than the parties themselves (Polinsky 1980; Krier and Schwab 1995). The common law, it should be noted, tends to prefer injunctions more as damages are more difficult for an external observer to calculate. For example, breach of an agreement to sell land, thought to be unique, is usually remedied by specific performance. However, breach of an agreement to sell a commodity is generally remedied by expectancy damages. In 1972 then professor Richard A. Posner analyzed these and many other questions in his regularly updated book *Economic Analysis of Law* (Posner 1972), which was explicitly indebted to Coase and in a real sense institutionalized law and economics in legal analysis.

A "Coasean market" is one in which all affected parties must agree before a particular transaction can occur. In a traditional neoclassical market, by contrast, there might be thousands of buyers and sellers but only one of each is necessary to a deal. For example, when one buyer purchases bread from one seller, the rest of the market does not participate and is largely indifferent. The difference among markets is readily apparent in a case such as Sturges vs. Bridgeman. While Victorian London contained thousands of physicians, confectioners, and suitable houses, the "market" that Coase discusses involved a single seller, a single buyer, and a single duplex house. In the long run either Sturges or Bridgeman could avoid the conflict by moving way, thus indicating that Coase's focus is not merely on a very tiny market but also on the short run. This small grouping is a market to the extent that the costs of exiting exceed the costs of reaching a bargain and staying. One important impact of the Coase Theorem was to increase economists' focus on very small markets, such as the two parties to a tort dispute, a few homeowners in a subdivision, a husband and wife, or the relation between shareholders, creditors, and managers in a single firm.

The requirement that all parties in a Coasean market must agree poses difficulties as the number of parties increases or their interests are more diverse. For example, a smokestack factory might willingly compensate 100 downwind landowners in order to keep running. But each one may be entitled to an injunction (abatement of the nuisance), so all must agree about how to share the award. More adjacent landowners, larger landowners, or those with more valuable homes will seek a larger share, and until these issues are resolved, there will be no agreement. The result could be endless cycles of coalitions and countercoalitions. That this is a consequence of high transaction costs is by no means clear. A rational participant bargains as long as the cost of a further offer is less than the expected payoff. So if bargaining were indeed costless but there was any uncertainty about outcomes, bargaining would not stop. In these situations positive transaction costs force the agreement by making continuing bargaining more costly at the margin than any expected payoff. The transaction/bargain cost curve thus has a lopsided "U" shape, with endless bargaining when transaction costs are near zero, more successful bargaining when they are a little higher, and less successful as they rise to yet higher levels.

When multiparty Coasean bargains do occur, they can result in excessive stability. Exiting from them can be just as difficult as entering them in the first place. For example, zoning laws can usually be changed by the majority vote of a legislative body as economic conditions change. By contrast, contractual servitudes generally require the unanimous consent of all affected parties, producing significant holdup problems when the majority believes or market values indicate that a servitude has become counterproductive (Hovenkamp 2002).

Coase acknowledged many of these difficulties in his 1959 article on the Federal Communications Commission, although he paid little attention to them later and much of the transaction cost literature has ignored them. Coase noted that "when large numbers of people are involved, the argument for the institution of property rights is weakened and that for general regulations becomes stronger." Speaking of smoke pollution, he acknowledged that "if many people are harmed and there are several sources of pollution, it is more difficult to reach a satisfactory solution through the market." As a result, "in these circumstances it may be preferable to impose special regulations. . ."(Coase 1959, at 27, 29).

These admissions invite the question whether Coase really attacked Pigou fairly. The argument for Pigouvian taxes was not concerned about conflicting rights as between two bargainers where no one else was affected. Rather, it was with problems such as highway congestion or pollution, which affect many users, both spatially and often temporally. As a result Pigouvian taxes, such as a carbon tax, continue to have support among mainstream economists (e.g., Mankiw 2012, pp. 207–210; Medema 2011; Baumol 1972). The cost of fossil fuels includes not only production and distribution costs but also longer-run environmental costs. The affected interests include hundreds of millions of people and even future generations.

In the 1960s and 1970s, "The Nature of the Firm" was rediscovered. Together with "The Problem of Social Cost," they became cornerstones in the development of "New Institutional Economics" (NIE) and its variations, sometimes called "organizational economics" or "transaction cost economics." The earlier work of Oliver E. Williamson, particularly *Markets and Hierarchies*, probably did more than anything to bring "The Nature of the Firm" into the spotlight (Williamson 1975). NIE refocused economic study on "institutions," an idea developed by the first generation of institutionalist economists a half century earlier, including Thorstein Veblen, Richard T. Ely, and John Commons. But NIE was dramatically different from the generally nontechnical, evolutionary, and often anti-marginalist conceptions of the original institutionalists (Hovenkamp 2013). The general thrust of NIE was to move economics away from large traditional markets to the study of very small ones, even viewing relationships inside the organization as a market. While the Coasean literature as applied to separate economic actors spoke of "transaction costs" as interfering with the efficient allocation of resources, the concept of "agency costs" came to describe costs internal to the firm that might obstruct efficient value maximization (e.g., Jensen and Meckling 1976). Another result was more refined studies of the risk and cost profiles that firms faced in deciding whether and how to integrate, including the significant costs of making costly, specialized commitments to one's trading partner (e.g., Klein et al. 1978; Coase 2000).

Coase made several other contributions to economics, including law and economics. One was his 1946 article "The Marginal Cost Controversy." In 1938 Harold Hotelling had argued that, because marginal cost pricing is essential to competition, outcomes in industries with very high fixed costs, such as railroads and electric utilities, would be suboptimal (Hotelling 1938). Prices would be driven to marginal cost, with insufficient surplus to cover fixed costs. The correction was government subsidies permitting such firms to recoup their fixed cost investment. Coase's rejoinder was to develop the concept of two-part pricing, with an entry or access fee to cover the fixed cost component and a per use variable fee to cover the marginal component (Coase 1946). Most of the subsequent literature on Coase's argument concludes two things. *First*, two-part pricing will rarely yield optimal outcomes when competitive providers set their own prices, although they are often more efficient than purely linear pricing. Second, however, two-part tariffs can be (and are) an effective way to encourage closer-to-optimal output in price-regulated markets by bringing the per use price closer to the marginal cost (Tirole 1988, pp. 142–146; Brown et al. 1992).

Two of Coase's important contributions in the early 1970s concerned the diverse topics of the durable goods monopolist and the scope of public goods. An article on durability and monopoly developed the "Coase conjecture" that in the very act of selling, the monopolist of a durable good dissipates its monopoly power (Coase 1972). It ends up competing with its own previous output, resold on secondary markets. Durability varies considerably, from nearly perfect in the case of land (Coase's opening example) to highly imperfect in the case of clothing (a used suit at Goodwill is a poor competitor for a new suit at Macys). The value of such a monopolist's output, Coase argued, depends on its ability to make a credible commitment to limit future output. For example, while van Gogh's painting *The Starry Night* (1889) is priceless, people can obtain a copy for $5 because the painting is in the public domain and no one can make a credible commitment that future output will be limited. In

addition, the durable goods monopolist may be able to profit by leasing rather than selling.

In "The Lighthouse in Economics" (Coase 1974), Coase wondered about the extent to which traditionally defined public goods really constitute a market failure. The lighthouse had appeared frequently in the economics literature as a public good that needed to be supplied by the government. However, Coase observed, privately owned lighthouses existed and were typically supported by a harbor tax or equivalent assessment against the vessels that benefitted from them. The real problem lays in developing an appropriate pricing mechanism and accounting for free riders – in particular, ships that might pass without actually using the harbor, thus benefitting from the lighthouse without paying the tax. Later critics observed, however, that private lighthouses either did not exist at all or else were short-lived relatively unsuccessful ventures (Bertrand 2006; Barnett and Block 2007). The harbor tax, if assessed by a public authority, was Pigouvian in any event.

Coase revisited many of the themes that defined his career in his Nobel Prize Lecture in 1991, entitled "The Institutional Structure of Production" (Coase 1992). He reiterated the theme that transaction costs are what give the legal system its importance and called for further empirical study of the role of transaction costs in real-world economies. He also lamented that his theories had been much less influential among economists than among lawyers – a view that was largely undermined by Oliver Williamson's receipt of the Noble Prize in 2009. Coase's work remains as alive and controversial as ever and has cast a long shadow on the disciplines of both economics and law.

References

Bain JS (1959) Industrial organization. Wiley, New York

Barnett W, Block W (2007) Coase and Van Zandt on lighthouses. Public Financ Rev 35:710–733

Baumol WJ (1972) On taxation and the control of externalities. Am Econ Rev 62:307–322

Bertrand E (2006) The Coasean analysis of lighthouse financing: myths and realities. Camb J Econ 30:389–402

Blaug M (2007) The fundamental theorems of modern welfare economics, historically contemplated. Hist Polit Econ 39(2):185–207

Blum WJ, Kalven H (1964) Public law perspectives on a private law problem–auto compensation plans. U Chi L Rev 31:641–723

Boulding KE (1942) The theory of the firm in the last ten years. Am Econ Rev 32:791–802

Brown DJ, Heller WP, Starr RM (1992) Two-part marginal cost pricing equilibria: existence and efficiency. J Econ Theory 57:52–72

Calabresi G (1970) The cost of accidents: a legal and economic analysis. Yale University Press, New Haven

Calabresi G, Melamed A (1972) Property rules, liability rules, and inalienability: one view of the cathedral. Harvard Law Rev 85:1089–1128

Cheung SNS (1983) The contractual nature of the firm. J Law Econ 26:1–21

Clark JM (1923) Studies in the economics of overhead costs. University of Chicago Press, Chicago

Coase RH (1937) The nature of the firm. Economica 4(n.s.):386–405

Coase RH (1946) The marginal cost controversy. Economica 13:169–182

Coase RH (1959) The federal communications commission. J Law Econ 2:1–40

Coase RH (1960) The problem of social cost. J Law Econ 3:1–44

Coase RH (1972) Durability and monopoly. J Law Econ 15:143–149

Coase RH (1974) The lighthouse in economics. J Law Econ 17:357–376

Coase RH (1988a) The nature of the firm: origin. J Law Econ Organ 4:3–17

Coase RH (1988b) The firm, the market and the law. University of Chicago Press, Chicago

Coase RH (1992) The institutional structure of production. Am Econ Rev 82:713–719

Coase RH (1994) Duncan Black. In: Ronald H (ed) Coase, essays on economics and economists. University of Chicago Press, Chicago, pp 187–189

Coase RH (2000) The acquisition of fisher body by General Motors. J Law Econ 43:15–32

Ellickson RC (1973) Alternatives to zoning: covenants, nuisance rules, and fines as land use controls. U Chi L Rev 40:681–781

Hotelling H (1938) The general welfare in relation to problems of taxation and of railway and utility rates. Econometrica 6:242–269

Hovenkamp H (1990) Marginal utility and the Coase Theorem. Cornell L Rev 75:783–801

Hovenkamp H (2002) Bargaining in Coasean markets: servitudes and alternative land use controls. J Corp L 27:519–530

Hovenkamp H (2011) Coase, institutionalism, and the origins of law and economics. Ind L J 86:499–542

Hovenkamp H (2013) The opening of American law: neoclassical legal thought, 1870–1970. Oxford University Press, New York

Jensen MC, Meckling WH (1976) Theory of the firm: managerial behavior, agency costs and ownership structure. J Fin Econ 3:305–360

Kahneman D et al (1990) Experimental tests of the endowment effect and the Coase Theorem. J Polit Econ 98:1325–1346
Klein B et al (1978) Vertical integration, appropriable rents, and the competitive contracting process. J Law Econ 21:297–326
Krier JE, Schwab SJ (1995) Property rules and liability rules: the cathedral in another light. New York U Law Rev 70:440–483
Mankiw NG (2012) Principles of economics, 6th edn. Cengage Learning, Independence, Ky
Marshall A (1890; 8th ed. 1920), Principles of economics. Macmillan, London
Medema SG (2011) Of Coase and carbon: The Coase theorem in environmental economics, 1960–1979 (SSRN working paper, Dec. 20, 2011), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1929086
Medema SG (2013) Rethinking market failure: 'The problem of social cost' before the 'Coase Theorem' (SSRN working paper, Jan. 25, 2013), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2188728
Pigou AC (1928) An analysis of supply. Econ J 38:238–257
Pigou AC (1932) The economics of welfare, 4th edn. Macmillan, London
Polinsky MA (1980) Resolving nuisance disputes: the simple economics of injunctive and damage remedies. Stanford Law Rev 32:1075–1112
Posner RA (1972, 8th ed. 2010) Economic analysis of law little. Little Brown/Aspen, Boston/New York
Robbins L (1928) The representative firm. Econ J 38:387–404
Robinson EAG (1931) The structure of competitive industry. Nisbet, London
Stigler GJ (1966) The theory of price, 3rd edn. Macmillan, New York
Stigler GJ (2003) Memoirs of an unregulated economist. University of Chicago Press, Chicago
Sturges v. Bridgeman (1879) 11 Ch. D 852
Tirole J (1988) The theory of industrial organization. MIT Press, Cambridge
Williamson OE (1975) Markets and hierarchies: analysis and antitrust implications. Free Press, New York

Coase and Property Rights

Elodie Bertrand
CNRS and University Paris 1 Panthéon-Sorbonne (ISJPS, UMR 8103), Paris, France

Abstract

In "The problem of social cost" (1960), Ronald H. Coase argued that what are exchanged are property rights and that the operation of the price system requires these rights to be defined. Coase was more interested in how property rights are (or should be) allocated and exchanged than in their content or definition. He insisted that factors of production must be considered as property rights, but conversely, property rights, even when they relate to nuisances, are nothing more than extra production costs.

Introduction

Ronald H. Coase was concerned with law since he studied at the London School of Economics. He was there greatly influenced by his professor then colleague Arnold Plant, who wrote extensively on the importance of the delimitation of property rights and on the influence of their structure on the economic system (see Coase 1977). "The problem of social cost" (Coase 1960) was addressing economists, and one of its main insights was to explain "that what are traded on the market are not, as is often supposed by economists, physical entities, but the rights to perform certain actions, and the rights which individuals possess are established by the legal system" (Coase 1992, p. 717). In fact, Coase developed this idea while writing on the Federal Communications Commission (Coase 1959).

The Role of Property Rights

In 1959, Coase famously argued that the price system could be used to allocate radio frequencies. His strategy was to demonstrate that frequencies are unspecific compared to other goods and services. First, just like a land must be possessed to avoid multiple uses of it, property rights on frequencies must be defined to avoid simultaneous uses of the same frequencies: "A private-enterprise system cannot function properly unless property rights are created in resources, and, when this is done, someone wishing to use a resource has to pay the owner to obtain it. Chaos disappears; and so does the government except that a

legal system to define property rights and to arbitrate disputes is, of course, necessary" (Coase 1959, p. 14). Coase was thus stressing that what are exchanged are property rights (like John Roger Commons before him) and that the operation of the price system requires these rights to be defined.

Second, Coase added that interferences between adjacent frequencies did not make property rights on frequencies specific either. Based on the Sturges v. Bridgman case (1879), he argues that either the confectioner has the right to make noise and imposes costs on his doctor-neighbor, who cannot longer practice, or the doctor has the right to practice in silence and he imposes costs on his neighbor. This is exactly the same with the owner of a land who impedes others to use it: "What this example shows is that there is no analytical difference between the right to use a resource without direct harm to others and the right to conduct operations in such a way as to produce direct harm to others. In each case something is denied to others: in one case, use of a resource; in the other, use of a mode of operation" (ibid.: 26). Likewise, this is the right to use frequency in a certain way that is exchanged, not the frequency itself: "What does not seem to have been understood is that what is being allocated by the Federal Communications Commission, or, if there were a market, what would be sold, is the right to use a piece of equipment to transmit signals in a particular way. Once the question is looked at in this way, it is unnecessary to think in terms of ownership of frequencies or the ether" (ibid.: 33).

This "change of approach" (Coase 1960, p. 42) is detailed in "The problem of social cost". Factors of production (including those that create external effects) must be thought of as property rights (and conversely): "A final reason for the failure to develop a theory adequate to handle the problem of harmful effects stems from a faulty concept of a factor of production. This is usually thought of as a physical entity which the businessman acquires and uses (an acre of land, a ton of fertiliser) instead of as a right to perform certain (physical) actions. . . . If factors of production are thought of as rights, it becomes easier to understand that the right to do something which has a harmful effect (such as the creation of smoke, noise, smells, etc.) is also a factor of production. . . . The cost of exercising a right (of using a factor of production) is always the loss which is suffered elsewhere in consequence of the exercise of that right" (ibid.: 43–44). Nuisances are defined as reciprocal conflicts over the use of a property right. Like for other factors of production, the cost of this right is an opportunity cost, and the price system makes the exchanges efficient, if it operates without cost.

The role of the judge would then just be to define property rights, no matter how, but in a definite and predictable way (ibid.: 19). Exchanges on these property rights (including those whose use implies effects on others) could then take place and yield an optimal result, independent from their initial allocation: this is the idea Stigler named "the Coase theorem".

However, transaction costs may prevent some exchanges of rights, and, when this is the case, the initial allocation of rights is not modified or, at least, not until the optimal allocation is reached. In these conditions, the initial distribution of property rights influences the economic result: "with positive transaction costs, the law plays a crucial role in determining how resources are used" (Coase 1988, p. 178). In this case, what should be done?

Either the initial delimitation of rights is given but inefficient, and the economist or the policy-maker must compare the values of production yielded by different institutional arrangements (market, firm, regulation, status quo) and choose the one in which it is the highest, taking into account the costs of operation of these arrangements and the costs of changing from one to another (Coase 1960, pp. 16–18).

Either property rights are not yet allocated and common law judges should take into account this economic influence of their decisions when allocating property rights: "It would therefore seem desirable that the courts should understand the economic consequences of their decisions and should, insofar as this is possible without creating too much uncertainty about the legal position itself, take these consequences into account when making their decisions. Even when it is

possible to change the legal delimitation of rights through market transactions, it is obviously desirable to reduce the need for such transactions and thus reduce the employment of resources in carrying them out" (ibid.: 19). This entails distributing, right from the start, the property right to the person who values it the most, that is to say, imitating the result of the market. If exchanges cannot take place, this could improve efficiency. Even when transaction costs do not prevent exchanges of the right, limiting the need for exchanges economizes on these costs. And Coase couples this normative role of judges to the empirical claim that they actually are, at least partly, aware of the reciprocity of the problem and of the economic consequences of their decisions: they introduce economic efficiency considerations in their deliberations, as several cases analyzed in "The problem of social cost" would suggest (see Bertrand 2015). This was the very beginning of the debate over the efficiency of the common law, more famously brought to the fore by Posner (1972).

Coase, however, also mentions that the allocation of property rights by statutory law (by contrast to common law) is generally inefficient, because it protects harmful producers – gives them the right to pollute – beyond what would be economically desirable (Coase 1960, pp. 26–27).

Regarding the analysis of property rights in law and economics, Coase introduced three important ideas. First, property rights may be analyzed with the theory of markets and contracts; this kind of work was developed by Armen Alchian, Harold Demsetz, and Douglass North. Second, externalities are the symptom of ill-defined property rights, as was substantiated by Demsetz (1967). Third, and this intuition was expanded by Douglas Allen (1991) and Yoram Barzel (1989), transaction costs are the costs of establishing and maintaining property rights.

What Are Coasean Property Rights?

Coase was more interested in how property rights are (or should be) allocated and exchanged than in the content or definition of these property rights. "The problem of social cost" uses the term "property right", but in the examples on which the analysis is based, agents are "liable" or not, and the reciprocal situations of liability are not exactly symmetrical.

Calabresi and Melamed's (1972) typology helps to reinterpret Coase's argument. When the rancher is "liable" (Coase's words), this means that he has to pay the farmer for the damage caused to his corn. The "entitlement" is here protected by a "liability rule" (Calabresi and Melamed's terms): the agreement of the farmer is not necessary for the exchange to take place, and the price is externally determined. In the reciprocal case, when the rancher is not "liable", the farmer has to negotiate with him so that he diminishes his herd; this means that this time the entitlement is protected by a "property rule": the rancher's agreement is necessary, and the price is determined during the process of negotiation.

Merrill and Smith (2001, 2011) explain that Coase thinks property rights as bundles of use rights and that he is the one who transmitted this legal realist view of property to (law and) economics. This view is very different from the traditional conception of property, attached to "things" and excluding the world from this "thing" (in rem property), that we find in William Blackstone and Adam Smith. Coase is rather referring to a bundle of in personam rights, attached to persons and obtained against certain other persons. This is most visible here: "We may speak of a person owning land and using it as a factor of production but what the land-owner in fact possesses is the right to carry out a circumscribed list of actions" (Coase 1960, p. 44). Coase chose this conception because he was confronted to radio frequencies which are not "things" (Coase 1988, p. 11): he was dealing with harmful effects and was concerned with exchanges of rights; but this choice obscured the in rem character of property rights.

Conclusions

Coase's view of property rights as bundle of rights comes from his will to conceive harmful effects as

unspecific by thinking of them as just another factor of production. As soon as 1959, Coase asserts that, in order to practice, the doctor must own not only his examination room but also the right to use it in silence; likewise the confectioner must own not only his machinery but also the right to use it noisily. The right to harm or to be protected from harms complements the classical factors of production: it is a right to use a certain resource in a certain manner, and it is a cost, to be taken into account among the other costs of production. This idea permitted considering the right to harm or to be protected from harms as a factor of production. But this is also how, while Coase wanted to introduce property rights in economics, he assimilated them to unspecific factors of production, allowing them to enter into the analysis just as other costs of production. In Coase's examples, property rights are just costs. He insisted that factors of production must be considered as property rights, but conversely, property rights, even when they relate to nuisances, are nothing else than extra production costs. We therefore understand how this vision of property rights disregards any notion of causality, responsibility, enmity, or moral inalienability and more largely any historical, social, or moral dimension.

Cross-References

- ▶ Coase, Ronald
- ▶ Coase Theorem
- ▶ Externalities
- ▶ Transaction Costs

References

Allen DW (1991) What are transaction costs. Res Law Econ 14:1–18
Barzel Y (1989) Economic analysis of property rights. Cambridge University Press, Cambridge
Bertrand E (2015) 'The fugitive': the figure of the judge in Coase's economics. J Inst Econ. 11(2):413–435
Calabresi G, Melamed AD (1972) Property rules, liability rules, and inalienability: one view of the cathedral. Harv Law Rev 85:1089–1128
Coase RH (1959) The federal communications commission. J Law Econ 2:1–40
Coase RH (1960) The problem of social cost. J Law Econ 3:1–44
Coase RH (1977) Review of: selected economic essays and addresses by Arnold Plant. J Econ Lit 15:86–88
Coase RH (1988) The firm, the market and the law. The University of Chicago Press, Chicago
Coase RH (1992) The institutional structure of production. Am Econ Rev 82:713–719
Demsetz H (1967) Toward a theory of property rights. Am Econ Rev 57:347–359
Merrill TW, Smith HE (2001) What happened to property in law and economics? Yale Law J 111:357–398
Merrill TW, Smith HE (2011) Making Coasean property more Coasean. J Law Econ 54:S77–S104
Posner RA (1972) Economic analysis of law. Little, Brown, Boston

Coase Theorem

Steven G. Medema
Department of Economics, University of Colorado Denver, Denver, CO, USA

Definition

Assuming the property rights are well defined and that the costs of transacting are zero, parties to an externality will resolve the dispute efficiently, and the outcome will be unaffected by to which party rights are initially assigned.

Introduction

The Coase theorem was derived from the negotiation result laid out by Ronald Coase in his 1960 article, "The Problem of Social Cost" (1960), after having first been articulated in his discussion of the allocation of broadcast frequencies a year earlier (Coase 1959). The theorem, so named by George Stigler (1966, p. 113), has been stated in a variety of ways by the thousands of authors who have invoked it over the last five decades, but the essentials are as follows: Assuming the property rights are well defined and that the costs of transacting are zero, parties to an externality will resolve the dispute efficiently, and the outcome will be unaffected by to which party rights are

initially assigned. In short, rights matter, but to whom they are assigned initially does not. The theorem thus suggests that, under assumed conditions, the assignment of rights related to pollution externalities or of liability for accidents resulting from the use of (perhaps defective) consumer products, the remedy mandated for breach of contract, the rules governing trespass, etc. are largely irrelevant. It matters crucially that there are legal rules to govern these phenomena, but to whom the relevant rights are assigned does not. An efficient and invariant allocation is guaranteed regardless.

The Coase theorem is a cornerstone of the economic theory of externalities and of the economic analysis of law; yet, it remains the subject of controversy. It has been the subject of challenges and defenses both intuitive and mathematical, experimented with and subjected to empirical assessment. (Medema and Zerbe (2000) provide a survey that includes a plethora of references to the literature discussed in the present essay.) There are many who believe that the theorem is not valid as a proposition in economic logic, but the theorem is nonetheless referenced regularly as an accepted truth in the scholarly and textbook literatures and has been applied in various ways across virtually every subfield of economics and of law, to the political realm, and beyond. Where their *is* widespread agreement is on the theorem's domain of *direct* relevance – that it is highly circumscribed and likely zero. At issue here is the theorem's assumption of zero transaction costs, a frictionless world that, one could argue, is more highly restrictive than the world of perfect competition – the economist's ideal type.

Underpinnings

The zero transaction costs assumption remains imprecisely defined even to this day owing to the ambiguity surrounding the notion of transaction costs. Meanings attributed to the zero transaction costs assumption range from the minimalist idea that there are no costs associated with engaging in negotiations per se (in essence, that talk is free) to the vastly more expansive idea that all relevant information can be acquired costlessly and thus that mutual gains can be realized instantaneously (Allen 1990). One gets a sense at times that, in the hands of those supporting the theorem, the working definition of transaction costs is "all impediments to the Coase theorem's validity." While this can give the Coase theorem a tautological air, the fact is that the frictionless world of the theorem has its counterparts in physics and elsewhere, and its unrealistic nature does not leave it without significance as a framework for thought experiments. More troubling on this score is that the Coase theorem assumes a free lunch that, at a minimum, there are no opportunity costs of time; the only question is how vast this free lunch domain is assumed to be.

Many of the critiques that are said to invalidate the Coase theorem involve violations of the zero transaction costs assumption, or at least can be objected to on those grounds. This is particularly true of the critiques grounded in game theory, the strategic behavior accompanying which is only possible because of some form of imperfect/incomplete information. Thus, while there is no question – as has been demonstrated time and again – that game theoretic analysis reveals very clearly that there are any number of settings in which the Coase theorem's efficiency and invariance predictions do not hold up, there is substantial question as to whether the conclusions drawn actually go to the validity of the Coase theorem itself.

The second fundamental assumption underlying the theorem is that property rights – that is, who has the right to do what – are fully specified over the relevant resources. These rights make clear the baseline against which any negotiations will take place and precisely what it is that is being exchanged between parties through any negotiation process. Absent such well-defined rights, policing and enforcement costs may be sufficiently high to preclude negotiations entirely or, at the very least, inhibit the exploitation of all relevant gains from exchange (Demsetz 1964). Some have argued that the property rights assumption is subsumed under or simply an extension of the assumption of zero transaction costs (or at least the most broad version of it), as, within such a world, there will be no uncertainties about the status of rights and no costs of policing and enforcement. Of course, these rights must be

alienable, for otherwise the associated costs of transaction are effectively infinite.

The Consumer Problem

When Coase laid out his negotiation analysis in "The Problem of Social Cost," both his analysis and the illustrations that he employed dealt with externality relationships among business firms. While, as any number of subsequent commentators pointed out, this presents its own set of potential problems – e.g., entry-related long-run asymmetries, non-separable cost functions, non-convexities in production sets, and the like – these objections can be overcome with relative ease (though validity here, like beauty, is sometimes in the eye of the beholder). Where the Coase theorem has run into difficulty is when consumers are party to the externality, whether as victims of, say, pollution emitted by a factory or when the externality is among to individual agents, such as when a neighbor plays his/her stereo too loud.

The challenges for the theorem pointed to here take two forms, one of which has been present in the literature since the mid-1960s and the other of which has come to the fore more recently via scholarship in the area of behavioral economics. The first of these is the problem of income effects, that the initial assignment of rights raises the income/wealth of one party relative to the other (Mishan 1971). The different patterns of resulting demands can both give rise to varying negotiated solutions to the externality issue and, in a large numbers situation, varying levels of prices and outputs in the marketplace. Thus, while the negotiated solution may be efficient (in the sense that all gains from trade have been exhausted), the outcome is not invariant – or even, strictly speaking, comparable – across alternative specifications of rights. One result of this challenge has been a propensity to add an "income effects aside" qualification to statements of the theorem or statements of the theorem that include the efficiency result but not the invariance claim – though these are far from universal in the literature.

The second issue on the consumer side arises when the initial distribution of rights impacts the valuation that agents place on the rights in question (Kahneman et al. 1990). The problem arises because of the potential that a given right may be valued differently by A when he/she owns that right than when that same right is owned by B. The potential for a divergence between an individual's willingness to pay for a right (WTP) and the amount which he/she is willing to accept in payment (WTA) to give up that right holds out the prospect that outcomes of the exchange process will vary if WTA $\neq$ WTP and, in fact, that such divergences may preclude exchange altogether if rights are assigned to one party rather than the other. The fact that a number of experimental treatments of the Coase theorem and of other economic contexts have provided evidence for the reality of such divergences has further called into question the validity of the theorem's invariance result when consumers are party to the dispute in question.

Why Does This Matter?

If the Coase theorem's validity depends on the existence of a fictional world many steps removed from reality and even then might not hold water, why has it been the subject of so much controversy and, even with that, come to occupy such a prominent place in legal-economic analysis?

First, the received theory of externalities at the time when Coase formulated his negotiation result implicitly assumed a similarly fictional world. In fact, even today the textbook analysis of externalities and of Pigovian remedies for them carries on that framework. The idea that Pigovian instruments – taxes, subsidies, and regulations – can be used to achieve efficient resolutions of externalities assumes that there are no costs, information, or otherwise associated with coordination. But the Coase theorem shows that, if coordination is costless, exchange, too, can generate efficient solutions to externality issues. That is, the claim that Pigovian instruments are *necessary* for the efficient resolution of externalities is shown to be invalid. Moreover, if one is going to invoke the reality of transaction costs against the Coase theorem, consistency requires taking into account the costs associated with governmental coordination. The implication of this is that there is no globally optimal means of dealing with externalities; the appropriate response (which may be

allowing the status quo to persist) can only be determined on a case-by-case basis.

Second, the Coase theorem shows that people will, of their own volition, efficiently and invariantly resolve externality issues if they are free to do so – that is, if transaction costs do not get in the way. This result, then, is the one that would be preferred by both parties against all other feasible alternatives. Given this, the argument can be (and has been) made that judges should assign rights so as to facilitate this outcome – that is, should "mimic the market" – so that the results which agents would arrive at in a frictionless world can be realized in the world in which we live. Differently put, the Coase theorem provides the philosophical foundation for the idea of efficiency as justice.

Third, the Coase theorem suggests that private agreements can be utilized to resolve externality issues within an appropriate legal framework – not that the theorem's predictions of an efficient and invariant result carry through in reality, but that significant efficiency gains can be realized if agents are placed within a context where rights are well defined and transaction costs are reasonably low. Such arrangements, it is argued, may in many instances (especially with small numbers of agents) prove to be superior, in an efficiency sense, to those arrived at via more traditional externality remedies. Thus, by demonstrating the veracity of markets/exchange in a frictionless world, the theorem suggests that its underlying processes may offer the best means of dealing with externalities in contexts not too far removed from its basic assumptions. Phenomena as disparate as out-of-court legal settlements, property rights solutions to common pool problems, and marketable emissions permits are seen as evidence of this implication of the Coase theorem's insights.

The Reciprocity Issue

The Coase theorem literature has had associated with it, almost from the start and in both economics and law, a ideological cast – a cast that has to do in part with divergent views regarding the relative efficacy of market and governmental coordination but also has at least as much to do with the question of the appropriateness of assigning rights to parties said, according to the conventional wisdom, to be the cause of the externality rather than the ostensible victims of it. But here, too, the Coase theorem set conventional wisdom on its head.

Given that the Coase theorem was born and came of age during a period of great concern over industrial pollution, its suggestion that efficiency will obtain regardless of to which party rights are initially assigned raised some disquiet, as the specter of victims being required to bribe polluters to achieve a reduction in pollution emissions raised its head. Similar concerns arose on the legal front, as scholars raised the possibility that victims would be required to bribe criminals to avoid being robbed or potential victims of tortious harm having to bribe potential injurers into taking precaution against, as in the case of products liability, acting recklessly or producing dangerous products.

The conclusion that the assignment of rights does not impact the allocation of resources reflects perhaps the most important insight provided by the Coase theorem: the reciprocal nature of externalities. This reciprocity can be conceptualized in two ways. The first goes to the notion of "causation," where the reciprocity issue informs us that it is improper to label one party as *the cause* of the harm. The agent typically labeled the "victim" of the factory's pollution is as much the cause of the harm as the factory itself: Take away either factory or neighbor, and the externality disappears. The second arena of reciprocity is on the costs front, as each party can be conceived of as a source of uncompensated costs imposed on the other. If the factory has the right to pollute, it visits uncompensated costs upon its neighbor by virtue of the pollution generated. If, on the other hand, the neighbor has the right to be free from pollution, costs are imposed upon the factory which must install pollution abatement equipment, move locations, cease production, or compensate the neighbor for harm caused. In sum, the imposition of costs/harm runs in both directions.

The theorem informs us that the externality will be resolved in efficient and invariant fashion regardless of to which party the relevant rights are initially assigned. But as we move away from the

theorem's idealized world of zero transaction costs, it also suggests that the assignment of rights and the form that these rights take has (perhaps significant) allocative import. Assignments of rights in one direction or another may do more to facilitate bargaining, owing to asymmetries in transaction costs, meaning that judicial decisions have the potential to either encourage negotiation or foreclose it. And when transaction costs are prohibitive, the utilization of liability rules rather than property rules may be expected to generate more efficient outcomes in the end. More generally, the theorem's emphasis on the reciprocal nature of externalities points us to the conclusion that the most efficient resolution of the problem may not be to restrain the actions of the party traditionally identified as the "cause" of the harm.

Conclusion

Questions of the theorem's validity notwithstanding, there can be no question that its influence in economics and in law has been considerable, even if the prolonged debate over it has generated more heat than light. The theorem suggested to economists the possibility of utilizing exchange or market processes to deal with real-world instances of externality where these possibilities had not previously been contemplated. In the legal realm, as well as the economic, the theorem emphasized that traditional notions of causality can be impediments to efficiency and that judicial decisions are not always the ultimate arbiter of rights. Perhaps most importantly, though, the frictionless world contemplated by the theorem brought to the fore the role played by transaction costs in market and exchange processes and the need to come to grips with the influence that legal and other institutions have on the magnitude of these costs.

Cross-References

- ▶ Coase and Property Rights
- ▶ Coase, Ronald
- ▶ Development and Property Rights
- ▶ Efficiency

References

Allen DW (1990) What are transaction costs? Res Law Econ 14:1–18
Coase RH (1959) The federal communications commission. J Law Econ 2(1):1–40
Coase RH (1960) The problem of social cost. J Law Econ 3:1–44
Demsetz H (1964) The exchange and enforcement of property rights. J Law Econ 7:11–26
Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the Coase theorem. J Pol Econ 98(6):1325–1348
Medema SG, Zerbe RO Jr (2000) The Coase theorem. In: Boudewijn BBA, Geest GD (eds) The encyclopedia of law and economics. Edward Elgar, Aldershot, pp 836–892
Mishan EJ (1971) The postwar literature on externalities: an interpretive essay. J Econ Lit 9(1):1–28
Stigler GJ (1966) The theory of price. Macmillan, New York

Further Reading

Ayres I, Talley E (1995) Solomonic bargaining: dividing a legal entitlement to facilitate coasean trade. Yale Law J 104(5):1027–1117
Buchanan JM (1973) The Coase theorem and the theory of the state. Nat Resour J 13:579–594
Calabresi G (1991) The pointlessness of pareto: carrying Coase further. Yale Law J 100:1211–1237
Calabresi G, Melamed AD (1972) Property rules, liability rules and inalienability: one view of the cathedral. Harv Law Rev 85(6):1089–1128
Coase RH (1988) Notes on the problem of social cost. In: The firm, the market, and the law. University of Chicago Press, Chicago, pp 157–185
Cooter R (1982) The cost of Coase. J Leg Stud 11(1):1–33
Demsetz H (1972) When does the rule of liability matter? J Leg Stud 1(1):13–28
Ellickson RC (1986) Of Coase and cattle: dispute resolution among neighbors in Shasta County. Stanford Law Rev 38(3):623–687
Farber DA (1997) Parody lost/pragmatism regained: the ironic history of the Coase theorem. Va Law Rev 83:397
Halteman J (2005) Externalities and the Coase theorem: a diagrammatic presentation. J Econ Educ 36(4):385–390
Hoffman E, Spitzer ML (1982) The Coase theorem: some experimental tests. J Law Econ 25(1):73–98
Medema SG (1994) Ronald H Coase. Macmillan, London
Medema SG (1999) Legal fiction: the place of the Coase theorem in law and economics. Econ Philos 15(2): 209–233
Medema SG (2009) The hesitant hand: taming self-interest in the history of economic ideas. Princeton University Press, Princeton
Parisi F (2003) Political Coase theorem. Public Choice 115(1/2):1–36

Posner RA, Parisi F (2013) The Coase theorem, vol 2. Edward Elgar, Cheltenham
Samuels WJ (1974) The Coase theorem and the study of law and economics. Nat Resour J 14:1–33
Usher D (1998) The Coase theorem is tautological, incoherent or wrong. Econ Lett 61(1):3–11

Coase Theorem and the Theory of the Core, The

Varouj A. Aivazian[1] and Jeffrey L. Callen[2]
[1]Department of Economics, University of Toronto, Toronto, ON, Canada
[2]Rotman School of Management, University of Toronto, Toronto, ON, Canada

Abstract

Aivazian and Callen (1981) and a number of their subsequent papers use cooperative game theory and core theory to show that the Coasean efficiency result is not robust when there are more than two players. Drawing primarily on their results, this chapter systematically explains the main argument and its extensions as follows. First, the Coase theorem could break down when there are more than two participants because the core of the negotiations may be empty under one set of property rights and nonempty under another. Second, transaction costs will tend to aggravate the empty core problem and make it more likely that the Coasean efficiency result will fail. Third, Pareto optimality can be achieved when the core is empty by the imposition of constraints on the bargaining process and the use of penalty clauses and binding contracts. Overall, the results indicate that it is important to distinguish between transaction costs (when the core exists) and costs due to the empty core because each has different implications for rationalizing institutions. This chapter also summarizes experimental results indicating that the existence of the core is an important determinant of negotiations generally and the Coase theorem in particular. It also points out that some of the problems raised for Coasean efficiency by the empty core also arise under alternative (non-core) notions of coalitional stability.

Introduction

The Coase theorem states that with well-defined property rights and in the absence of transaction costs, Pareto-efficient allocations will emerge through negotiations among the players to internalize any externality among them, regardless of the initial assignment of property rights (Coase 1960). This result obtains because participants will costlessly recontract around property rights assignments that fail to be Pareto efficient. Coase (1960) also emphasizes the central importance of transaction costs for resource allocation, focusing on efficient property right structures when transaction costs are significant.

As argued by a number of scholars, the bargaining mechanism over property rights in the Coase theorem can be fruitfully framed in terms of cooperative game theory (Arrow 1979; Davis and Whinston 1965). Focusing on core theory, a branch of cooperative game theory, the Coase theorem can be interpreted as: with zero transaction costs, the grand coalition will always emerge regardless of the initial allocation of property rights among the players, and irrespective of whether or not the core of a superadditive characteristic function exists. (Telser (1994) provides a compelling discussion of the power of core theory). Aivazian and Callen (1981) and a number of their subsequent papers employ cooperative game theory and core theory to show that while the Coasean efficiency result is robust for the case of a two-person game, it may fail when there are more than two players. We review these results drawing on Aivazian and Callen (1981, 2003), Aivazian et al. (1987), and Aivazian et al. (2009) as well as on some of the papers that commented on the original 1981 Aivazian-Callen study.

Aivazian and Callen (1981) show that the Coasean efficiency result may fail in a zero transaction cost environment with at least three players and two externalities, in which there are gains from cooperating and forming coalitions to

internalize the externalities. Specifically, in their example, the core is empty under one set of property rights, but nonempty under the other. With an empty core, cycling among coalitions could occur, preventing attainment of the grand coalition and Pareto efficiency. In response to the Aivazian and Callen counterexample to his theorem, Coase (1981) asserts that the zero transaction cost environment underlying his theorem is uninteresting in and of itself. He argues that if transaction costs are imposed on the negotiations in the Aivazian-Callen example, an empty core is less likely to obtain implying that the counterexample is essentially uninformative. However, extending their counterexample to allow for a reasonable transaction cost technology that is convex in the number of coalition partners, Aivazian and Callen (2003) demonstrate that transaction costs tend to aggravate the empty core problem making the breakdown of Coasean efficiency even more likely.

The Empty Core Argument without and with Transactions Costs

The original Aivazian and Callen (1981) analysis involves two polluting firms (A and B) and a laundry (C) and shows that when the polluting firms are liable (when the laundry has the property rights), the Pareto efficient outcome emerges; but when they are not liable, the core is empty and negotiations cycle without necessarily converging to the grand coalition or any other specific outcome. This can be demonstrated by representing the Aivazian and Callen (1981) example in the form of the following normalized characteristic function where V denotes joint coalitional profits:

$$V(i) = 0 \text{ all } i = A,B,C \tag{1a}$$

$$V(A,B) = a, V(A,C) = b, V(B,C) = c \tag{1b}$$

$$V(A,B,C) = d \tag{1c}$$

where a, b, c, d are positive constants, and $d > a$, b, c, for superadditivity. The Pareto optimal outcome corresponds to the grand coalition outcome V(A,B,C). Note that the characteristic function will be different under different property rights since what each coalition can guarantee itself depends on the prevailing property rights arrangements (Shubik 1984, Chap. 19). Necessary and sufficient condition for the core to be empty (when A and B are not liable) is

$$d < 1/2(a + b + c). \tag{2}$$

If the latter inequality obtains, the grand coalition outcome is not guaranteed so that specific property rights matter for efficiency.

Aivazian and Callen (2003) extend their 1981 paper to allow for transaction costs in the negotiation process by making the reasonable assumption that the costs of forming a coalition are convex in the number of players in the coalition, that is, coalition formation costs increase at an increasing rate with the number of coalition partners. This is reasonable since the number of communication channels required to obtain agreement among coalition members is also convex in the number of members. Two important conclusions emerge. First, if the core is empty in the absence of transaction (coalition formation) costs, then it is necessarily empty with such costs. Second, even if the core is not empty in the absence of transaction costs, such costs could generate an empty core.

What Have We Learned from the Empty Core Argument?

Several lessons emerge from the original Aivazian and Callen (1981) paper and the literature it has spawned (see, e.g., Bernholz (1997), De Bornier (1986), Hurwicz (1995), and Mueller (2003)). First, the Coase theorem may break down when there are more than two participants because the core of the negotiations may be empty under one set of property rights and nonempty under another. As a consequence, even in the absence of other transactions costs, the empty core is likely to impose particular costs of its own. Specifically, cycling induced by the empty core will tend to increase bargaining costs, diminish the value of the exchange opportunity as the negotiation

process is prolonged, and the exchange is postponed, potentially keep at least one coalition unsatisfied and yield a non-Pareto allocation of resources if eventually the grand coalition does not form (Aivazian and Callen (1981, 2003); Shubik (1983), p. 150–151). In fact, Bernholz (1997) argues that the empty core in the Aivazian and Callen example is equivalent to cyclical social preferences. Second, the empty core problem arises as the number of participants increases only when additional participants bring in additional externalities (see Mueller (2003), De Bornier (1986), and the discussion in Aivazian and Callen (2003) on this point). Third, transaction costs may well aggravate the empty core problem, especially if they are incurred prior to the bargaining process (Aivazian and Callen 2003; Anderlini and Felli 2006). Fourth, Pareto optimality may be achieved when the core is empty if there are reputational (transaction) costs with the breaking of agreements (Guzzini and Palestrini (2009), or, from a normative perspective, by the imposition of constraints on the bargaining process (e.g., limiting negotiations to only certain sub-coalitions) and the use of penalty clauses and binding contracts (Bernholz 1997; Telser 1994; Aivazian and Callen 2003). As a consequence, it is important to distinguish between transaction costs (when the core exists) and costs due to the empty core because each has a different implication for rationalizing institutions. As Aivazian and Callen (2003, p. 291–292) emphasize:

> It is wrong to conclude, therefore, that once transaction costs are introduced, then the problem of the empty core disappears and a Pareto optimal solution obtains. Rather, in such circumstances negotiations may break down more quickly and which specific coalition structure (the grand coalition or a proper sub-coalition) obtains cannot be specified a priori. Even if transaction costs were to force an equilibrium, nothing insures that the equilibrium is Pareto optimal … It may seem difficult to distinguish empirically between institutional arrangements that arise because of the nonexistence of the core from those that arise from the transaction costs of bargaining when there is a core. After all, the nonexistence of the core will also manifest itself in transaction costs, through the opportunity cost of (negotiation) time. However, the fact that an empty core can arise in the absence of bargaining costs, although these costs exacerbate the empty core problem, means that the costs generated by an empty core are fundamentally different from the transactions costs of bargaining. Indeed, what is unique about the empty core is that, in addition to direct bargaining costs, it gives rise to costs such as the erosion of the value of the exchange opportunity as it is postponed or the possibility of settling down to a non-Pareto optimal coalition (a proper sub-coalition).

Non-core Coalitional Stability

Many of the issues raised by the empty core also arise under alternative (non-core) notions of coalitional stability. As Aivazian et al. (1987) argue, for the Coase theorem to obtain, the grand coalition must be stable and, moreover, no other coalition can be similarly stable because otherwise a Pareto optimal allocation cannot be guaranteed. Aivazian et al. (1987) extend the Aivazian-Callen (1981) example to Aumman and Maschler (1964) bargaining set notions of coalitional stability by showing that while a specific Pareto optimal allocation of resources obtains for one set of property rights, a non-Pareto optimal allocation may well obtain for another set of property rights that involves bargaining. Indeed, under one type of bargaining set stability, they find that every coalition but the grand coalition is stable, completely vitiating the Coase theorem.

Testing the Implications of the Empty Core

It is unlikely that archival data are available that would allow one to test the implications of the empty core problem for the Coase theorem. Instead, Aivazian et al. (2009) investigate the Coase theorem experimentally in a bargaining game in which the final allocation of payoffs differ in terms of whether the core exists and in the initial allocation of property rights among the players. The experimental results indicate that the existence of the core is an important determinant of negotiations generally and the Coase theorem in particular. They find that when the core is empty and property rights are ill defined, Coasean efficiency breaks down. In particular, the number

of non-Pareto optimal agreements and negotiation rounds with cycling are significantly larger when the core is empty than when it exists, particularly when property rights are ill defined.

Conclusion

The upshot of the empty core issue for the Coase Theorem can be summarized as follows (Aivazian and Callen 2003, p. 296):

> "In the real world opportunities for exchange are sometimes manifold and the bargaining strategies potentially complex. The Coase Theorem masks this reality by presupposing that exchange occurs between two parties . . . with more than two parties, and at least two externalities, coalitional behavior may predominate. In which case, under some property rights arrangements the core may not exist; as a result, the Coase Theorem may fail to hold. Transactions costs may well exacerbate the empty core problem. In such circumstances, specific property rights arrangements, and contractual schemes such as penalty clauses, binding contracts, and restrictions on the sequence of bargaining, may emerge to attenuate the problems engendered by the nonexistence of the core"

Cross-References

- ▶ Coase and Property Rights
- ▶ Coase, Ronald
- ▶ Coase Theorem
- ▶ Coase Theorem: Empirical Tests

References

Aivazian VA, Callen JL (1981) The Coase theorem and the empty core. J Law Econ 24(1):175–181. Reprinted in R.A. Posner and F. Parisi (eds.) The Coase Theorem. Economics Approaches to Law Series. 2013. Edward Elger Publishing.

Aivazian VA, Callen JL (2003) The core, transaction costs, and the Coase theorem. Constit Polit Econ 14(4): 287–299

Aivazian VA, Callen JL, Lipnowski I (1987) The Coase theorem and coalitional stability. Economica 54(216):517–520. Reprinted in R.A. Posner and F. Parisi (eds.) The Coase Theorem. Economics Approaches to Law Series. 2013. Edward Elger Publishing.

Aivazian VA, Callen JL, McCracken S (2009) Experimental tests of core theory and the Coase theorem: inefficiency and cycling. J Law Eco 52(4):745–759. Reprinted in R.A. Posner and F. Parisi (eds.) The Coase Theorem. Economics Approaches to Law Series. 2013. Edward Elger Publishing.

Anderlini L, Felli L (2006) Transactions costs and the robustness of the Coase theorem. Econ J 116(508): 223–245

Arrow KJ (1979) The property rights doctrine and demand revelation under incomplete information. In: Boskin MJ (ed) Economics and human welfare: essays in honor of Tibor Scitovsky. Academic Press, New York, pp 23–29

Aumann RJ, Maschler M (1964) The bargaining set for cooperative games. In: Dresher M, Shapley LS, Tucker AW (eds) Advances in game theory. Princeton University Press, Princeton, pp 443–476

Bernholz P (1997) Property rights, contracts, cyclical social preferences and the Coase theorem: a synthesis. Eur J Polit Econ 13:419–442

Coase R (1960) The problem of social cost. J Law Econ 3(1):1–44

Coase R (1981) The Coase theorem and the empty core: a comment. J Law Econ 24(1):183–187

Davis OA, Whinston AW (1965) Some notes on equating private and social cost. South Eco J 32(2):113–126

De Bornier JM (1986) The Coase theorem and the empty core: a reexamination. Int Rev Law Econ 6(2):265–271

Guzzini E, Palestrini A (2009) The empty core in the Coase theorem: a critical assessment. Econ Bull 29(4): 3095–3103

Hurwicz L (1995) What is the Coase theorem? Jpn World Econ 7:49–74

Mueller DC (2003) Public choice III. Cambridge University Press, New York

Shubik M (1983) Game theory in the social sciences: concepts and solutions, vol 1. M.I.T. Press, Cambridge, MA

Shubik M (1984) Game theory in the social sciences: a game theoretic approach to political economy, vol 2. M.I.T. Press, Cambridge, MA

Telser L (1994) The usefulness of core theory in economics. J Econ Perspect 8(2):151–164

Coase Theorem: Empirical Tests

Elodie Bertrand

CNRS and University Paris 1 Panthéon-Sorbonne (ISJPS, UMR 8103), Paris, France

Abstract

This entry examines three classic tests of the Coase theorem: two empirical studies on the

most famous examples of externalities – Coase's ranchers and farmers (Ellickson, Stanford Law Rev. 38(3):623–687, 1986) and Meade's bees (Cheung, J Law Econ 16(1):11–33, 1973) – as well as the seminal laboratory experiments of Hoffman and Spitzer (J Law Econ 25(1):73–98, 1982). I will insist on the difficulty of testing the theorem without measuring transaction costs, and on another obstacle to negotiation over and above these costs: a moral or social prohibition on exchange.

Introduction

In *The Problem of Social Cost*, Coase (1960) used examples to suggest that, in the presence of externalities, if transaction costs are nil and if property rights are clearly defined and allocated, agents bargain over rights and achieve an optimal output that is independent of the initial allocation of rights. This proposition was to be called the "Coase theorem" by Stigler (1966, p. 113).

The theorem has long been tested against the facts, whether already existing (empirical studies) or purposefully constructed (laboratory experiments). Strictly speaking, however, these studies do not in fact test the Coase theorem. The reasons for this are twofold: first, the conclusion of almost six decades of debate over the validity of the theorem is generally taken to be that all criticisms can be subsumed under the category of transaction costs and that this renders the "theorem" tautological (Medema and Zerbe 2000). Second, transaction costs never being nil, the theorem does not apply to the real world (most of Coase's seminal article actually examined the consequences of these costs). Regarding experiments, they reduce transaction costs to a minimum. As for empirical studies, they can only test a "generalized Coase theorem" (Bertrand 2011): if the gain from a transaction concerning a right (the use of which provokes side effects) is greater than its cost, then the transaction takes place. I will insist on the difficulty of testing such a proposition without measuring transaction costs, and on another obstacle to negotiation over and above these costs: a moral or social prohibition on exchange. For a detailed review of the tests of the Coase theorem, see Medema and Zerbe (2000).

Empirical Studies of the Coase Theorem with Externalities

Some empirical tests of the theorem have concerned pretrial settlements (Galanter 1983) and transactions on nuisances after trials (Farnsworth 1999), but the majority test the prediction of the absence of any effects engendered by a change in the law: notably with regard to share tenancy arrangements (Cheung 1969), rules governing divorce (Peters 1986), the effect of a payment of bonuses to unemployed workers when they obtain a job or to employers when they hire unemployed people (Donohue 1989), and generally with a very specific attention to sports (Hylan et al. 1996; Cymrot et al. 2001; Szymanski 2007). I will focus on two tests which concern well-known examples of externalities: Coase's cattle that destroy crops on neighboring land (Ellickson 1986), and Meade's bees that pollinate orchards while foraging for nectar (Cheung 1973). The cattle case was also examined by Vogel (1987), who analyzed the effects of the changes in trespass law in the different counties of California during the second half of the nineteenth century: the assignment of rights to farmers tended to increase farm output, contrary to the prediction of the theorem (this movement was studied over a longer period than that of Ellickson, which explains their different results). Hanley and Sumner (1995) also studied the resolution of externalities due to red deer in Scotland, and their conclusions are close to Ellickson's.

Ellickson (1986) intended to test the realism of the "Parable of the Farmer and the Rancher" (p. 624), the numerical example developed by Coase (1960) in which, on the basis of their formal entitlements, a cattle raiser and a crops farmer negotiate the size of the herd in exchange for a monetary payment, and arrive at the same size whatever the initial allocation of rights is. In 1982, therefore, Ellickson turned to Shasta County, in the north of California, where two legal regimes coexisted: most of the county was under the historical "open-range" regime, in

which a cattleman is not liable for trespass damages even when negligent, but one district had switched to a "closed-range" regime in 1973, where the rancher is always liable for damage even in the absence of negligence.

Although the allocation of resources (e.g., spaces respectively devoted to crops and cattle) is not affected by changes in liability law, the explanation does not turn on the monetary negotiations invoked by Coase: neighbors rather refer to social norms to resolve their disputes; and since these norms are independent from formal law, the resolution of incidents is as well.

In particular, the parties to a dispute refer to an informal rule according to which the cattle owner is liable for the actions of his animals. This goes against the formal rule of the main part of the county, which implies that "norms, not legal rules, are the basic sources of entitlements" (Ellickson 1986, p. 672). Other norms detail the manner of resolving an incident: most of the time, the victim downplays the incident, which will be quickly solved by an exchange of civilities, and neighbors expect reciprocity. In any case, the resolution of the conflict must be informal, without recourse to law or courts, and monetary compensation is forbidden: inhabitants "regard a monetary settlement as an arms-length transaction that symbolizes an unneighborly relationship" (ibid.: 682).

This monograph therefore shows that neighbors do not solve their disputes with monetary transactions, but through social norms, which Ellickson explains by high transaction costs. But, first, he does not measure them: he simply infers them from the facts that exchanges do not take place and that the law would be costly to learn (a cost itself inferred from the fact that inhabitants do not know it). Second, it could be argued that transaction costs are low: small number of parties, easily identifiable, easy monetary assessment of the damages, sharing of the same social norms, etc. Third, other obstacles than transaction costs may explain the absence of exchange: social norms can impede a market from developing (Bertrand 2011). Take the norm that condemns monetary settlements: Ellickson explains this by reference to transaction costs, but we could just as well assert that it is this norm that prevents transactions in the first place, and hence that the norm lies at the origin of the impossibility of negotiations. Transaction costs would be irrelevant since agents are not willing to negotiate.

The absence of monetary payments and of reference to formal rules thus raises questions about the legitimization of the legal structure of rights and of their alienability, which may deter agents from negotiating these rights, quite apart from transaction costs. In the Shasta County case, there is no norm indicating that you can transact over the right to destroy your neighbor's crops. In fact, we here encounter the basic problem of externality: a market does not exist, and it may be difficult to create one out of nothing because of social norms. By contrast, the next example will stress that a market does seem to exist for what was long considered as externalities: pollination and nectar services.

Cheung's "Fable of the Bees" (1973) tests the realism of Meade's (1952) example of the externality between beekeepers and orchard owners. It was written at Coase's request, who was not satisfied with Johnson's study (1973) – the latter being more focused on the institutional setting of the pollination service market. Cheung's investigation was conducted in the state of Washington in spring 1972; it covered a sample of 9 beekeepers and a total of approximately 10,000 spring colonies.

Cheung first observes that a marketplace exists for transactions on pollination and nectar services. On the one hand, an orchard owner can rent the pollination services of an apiarist who places her hives in the orchard; he pays her a monetary pollination fee that depends on the number, density, and strength of hives. On the other hand, an apiarist who wants to place her hives among crops for honey production pays an apiary rent to the orchard owner, mostly in the form of honey, and depending on the honey crop. Note, however, a difference with Coase's parable of the rancher and the farmer: an orchard owner can choose a beekeeper from among others and vice versa; they are not compelled to negotiate or to find a solution with their neighbor.

Nevertheless, we here have evidence that contracts with monetary exchanges can deal with so-called externalities. It seems that the possibility of exchanges in pollination services (or social

permission) developed in the period after WWI, alongside with the specialization of farms which required specific pollinators. The right of having his plants pollinated was said to be exchangeable and a price was suggested: the set of social norms necessary for the operation of a market was available. Still, regarding the habit of giving some honey to the orchard owner in exchange for the right to place hives in his orchard, the question remains whether this should be considered as a market transaction or a social norm.

In any case, the markets for pollination and nectar services are unusual in that the enforcement of contracts, whether oral or written, relies heavily on social norms. In addition, externalities remain at the margins of the market for pollination services, and they are dealt with by social norms. For example, my neighbor, who also cultivates apples, benefits from the hives I rent for my apple trees. These positive externalities are solved by a social norm of neighborliness called "the rule of the orchards," by which we both rent the same number of hives per acre (Cheung 1973, p. 30). Cheung explains the absence of monetary exchange to solve this externality by the cost of such a negotiation (not measured, but itself inferred from the absence of such an exchange). As with trespass incidents, we could alternatively explain the absence of negotiation by the existence of an already satisfying norm or, prior to it, by other norms that would deter monetary negotiations between neighbors and make transaction costs irrelevant.

It would be excessive to infer from Ellickson's and Cheung's studies that the generalized Coase theorem is empirically confirmed, since both authors have a tendency to explain their observations precisely by this assumption. This kind of empirical study faces the problem of measuring the gains and costs of exchange. However, the control possible in the laboratory allows for a more precise determination of them.

Experiments of the Coase Theorem with Monetary Negotiations

Laboratory experiments concerning the theorem began in 1982 with the publication of studies by Prudencio (1982) and Hoffman and Spitzer (1982). It is the last protocol, closest to the Coasean parable of the rancher and the farmer, that has been used the most. Experiments of the theorem seek to test the robustness of results given variation of the implicit or explicit assumptions: high number of agents (Hoffman and Spitzer 1986); incomplete information (Prudencio 1982; McKelvey and Page 2000); introduction of transaction costs (time limit in Prudencio 1982 or cost of an offer in Rhoads and Shogren 1999); uncertainty about the payments (Shogren 1992); property rights earned and/or legitimized (Hoffman and Spitzer 1985a), not allocated (Harrison and McKee 1985) or uncertain (Cherry and Shogren 2005); non-convexity (Shogren et al. 2002); empty core (Aivazian et al. 2009); and physical discomfort (Coursey et al. 1987). Schwab's (1988) experiment does not concern externalities but rather contract presumptions. Another series of experiments consists in a comparison of the efficiency of different solutions to externalities (e.g., Plott 1983; Harrison et al. 1987). The results of the experiments on the endowment effect were applied to refute some of the theorem's predictions (see Kahneman et al. 1990) but are not specific to the theorem and do not exhibit theorem-like mechanisms.

It is Hoffman and Spitzer's (1982) first set of experiments, with two persons and complete information, whose design is the closest to the Coasean parable. The respective monetary payoffs (net profit) of the two subjects, who were randomly assigned the letters A (rancher) and B (farmer), depend on the value of a discrete number (the size of the herd): a couple of payoffs are associated to each number (0, 1, etc.) and only one number maximizes the sum of these payoffs. A's payoff increases when the number increases and conversely for B, which renders the "externality" negative. The allocation of the property right is translated by the designation of one of the subjects as the "controller": she has the right to unilaterally choose the number (and hence the payoffs). The possibility of negotiation comes from the fact that the other subject may influence her choice by proposing to transfer a part of his own payoff. The controller is randomly designated through a heads or tails game.

As for the results, exchanges take place and are efficient (23 out of 24 decisions choose the number that maximizes the sum of payoffs). The authors infer that this result, confirmed by their other sets (1982) and other experiments (1985a; 1986), "creates a strong presumption in favor of the Coase Theorem" (1985b, p. 1011). Nevertheless, most of the exchanges are not mutually advantageous. The controller sacrifices a part of her gain for fairness. Typically, whereas the controller, B, could obtain \$12 (and A 0) by unilaterally choosing the number 0, A and B sign an agreement by which the controller chooses the number 1 (B earns \$10, and A 4) – which is the maximum total payoff – and share this total gain equally (\$7 for each), which means the controller sacrifices \$5 (Hoffman and Spitzer 1982, p. 86 and 92). As Harrison and McKee (1985, p. 655) conclude, "the Coase Theorem is [therefore] behaviorally 'right for the wrong reasons.'" In these experiments, subjects agree on the optimal issue, but only because one of them sacrifices a part of her gains. Why does this subject accept the exchange?

Admittedly, fairness is a classic result of bargaining experiments, the robustness of which has been confirmed in large measure by Hoffman and Spitzer's (1982) own experiments. Fairness is here favored by public face-to-face negotiation; complete information (McKelvey and Page 2000 obtain less good results with incomplete information); and a random allocation of the right, without legitimacy and without insistence on its meaning (the controller obtains her individual maximum more often when the initial allocation of the right follows a preliminary game and when her authority is legitimated by the monitor (Hoffman and Spitzer 1985a), or when she can learn the meaning of her unilateral right (Harrison and McKee 1985)).

But fairness is here obtained by sacrifice. A likely explanation of this non-mutually advantageous exchange is that of a moral or social incentive raised by the experimental design and the instructions (Bertrand 2014). First, contracts are given to the subjects, from which they infer that the right is exchangeable and how they can exchange it. This amounts to morally authorizing the exchange regarding this externality, and even to encouraging it by implying that it is expected from participants that they use this possibility. Then, the list of payoffs in function of the number gives the subjects the monetary value of the externality, which impedes the perception of the moral problem of giving value to something that should not have one (Kelman 1985, pp. 1038–1039). Finally, instructions are not indifferent regarding the question of exchange, and take care, under the shelter of neutrality, to avoid any vocabulary concerning externalities, constraint, or causality.

Confronted with some of these criticisms, Coursey et al. (1987) built an experiment precisely to test the existence of a moral ban on bargaining in the presence of externalities. They replicate the 1982 experiment, but with the possibility of physical discomfort for the subject who suffers from the externality (keeping in his mouth a very bitter-tasting liquid for 20s). This allegedly allows testing the moral ban against negotiating over something that insults dignity, what they call the "dignity hypothesis." Pairs of subjects reached the optimal result, and authors hence reject this hypothesis. But here again, the sources of bargaining breakdown are in the most part avoided (Bertrand 2014, pp. 454–456). In particular, and paradoxically in comparison to what the authors wanted to test, the moral obstacles to the exchange are at least partially removed: the contract is provided, and the instructions and consent insist on the possibility of exchange and the possibility of A taking the liquid. Individuals placed in such a situation know what they have to do to please the monitor, or may simply internalize her authority (Milgram 1974).

Conclusion

Since the Coase theorem is circular, what is tested is precisely whether mutually advantageous bargains are indeed realized. Such realization may encounter three obstacles: (1) transaction costs, (2) problems of agreement over the distribution of the surplus, and (3) moral or social prohibition of exchange. This entry has shown that the authors of these tests commonly appeal to transaction costs to legitimize every result as efficient. They thereby rationalize their observations by appeal to

the assumption of efficiency rather than test its correspondence to the real world. And they underestimate the other obstacles.

What are the lessons for the theorem? As stressed by Medema, these studies enlighten us about "situational behavioral norms" (1997, p. 129) and the limits to the assumption of individual maximization, hence calling into question the behavioral assumptions that ground the theorem, and indeed the law and economics movement more generally. Cheung's bees and Hoffman and Spitzer's experiment bring to light that markets for what was seen as an "externality" can exist: the right over an "externality" is sold and bought. Two elements have nevertheless been overlooked by Coase (1960): first, the entitlements to which agents refer are not only determined by the law; second, impediments to bargaining other than transaction costs exist, these being moral or social obstacles to such an exchange.

Cross-References

- ▶ Coase Theorem
- ▶ Coase Theorem and the Theory of the Core, The
- ▶ Endowment Effect
- ▶ Nuisance

References

Aivazian VA, Callen JL, McCracken S (2009) Experimental tests of core theory and the Coase theorem: inefficiency and cycling. J Law Econ 52(4):745–759

Bertrand E (2011) What do cattle and bees tell us about the Coase theorem? Eur J Law Econ 31(1):39–62

Bertrand E (2014) Autorisation à l'échange sur des externalités. De l'interdiction à l'obligation. Revue Economique 65(2):439–459

Cherry TL, Shogren JF (2005) Costly Coasean bargaining and property right security. Environ Resour Econ 31(3):349–367

Cheung SNS (1969) The theory of share tenancy. University of Chicago Press, Chicago

Cheung SNS (1973) The fable of the bees: an economic investigation. J Law Econ 16(1):11–33

Coase RH (1960) The problem of social cost. J Law Econ 3:1–44

Coursey DL, Hoffman E, Spitzer ML (1987) Fear and loathing in the Coase theorem: experimental tests involving physical discomfort. J Leg Stud 16(1):217–248

Cymrot DJ, Dunlevy JA, Even WE (2001) "Who's on first": an empirical test of the Coase theorem in baseball. Appl Econ 33(5):593–603

Donohue JJ III (1989) Diverting the Coasean river: incentive schemes to reduce unemployment spells. Yale Law J 99(3):549–609

Ellickson RC (1986) Of Coase and cattle: dispute resolution among neighbors in Shasta county. Stanford Law Rev 38(3):623–687

Farnsworth W (1999) Do parties to nuisance cases bargain after judgment? A glimpse inside the cathedral. Univ Chicago Law Rev 66(2):373–436

Galanter M (1983) Reading the landscape of disputes: what we know and don't know (and think we don't know) about our allegedly contentious and litigious society. UCLA Law Rev 31:4–71

Hanley N, Sumner C (1995) Bargaining over common property resources: applying the Coase theorem to red deer in the Scottish Highlands. J Environ Manag 43(1):87–95

Harrison GW, McKee M (1985) Experimental evaluation of the Coase theorem. J Law Econ 28(3):653–670

Harrison GW, Hoffman E, Rutström EE, Spitzer ML (1987) Coasian solutions to the externality problem in experimental markets. Econ J 97(386):388–402

Hoffman E, Spitzer ML (1982) The Coase theorem: some experimental tests. J Law Econ 25(1):73–98

Hoffman E, Spitzer ML (1985a) Entitlements, rights, and fairness: an experimental examination of subjects' concepts of distributive justice. J Leg Stud 14(2):259–297

Hoffman E, Spitzer ML (1985b) Experimental law and economics: an introduction. Columbia Law Rev 85(5):991–1036

Hoffman E, Spitzer ML (1986) Experimental tests of the Coase theorem with large bargaining groups. J Leg Stud 15(1):149–171

Hylan TR, Lage MJ, Treglia M (1996) The Coase theorem, free agency, and major league baseball: a panel study of pitcher mobility from 1961 to 1992. South Econ J 62(4):1029–1042

Johnson DB (1973) Meade, bees, and externalities. J Law Econ 16(1):35–52

Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the Coase theorem. J Polit Econ 98(6):1325–1348

Kelman M (1985) Comment on Hoffman and Spitzer's "Experimental law and economics". Columbia Law Rev 85(5):1037–1047

McKelvey RD, Page T (2000) An experimental study of the effect of private information in the Coase theorem. Exp Econ 3:187–213

Meade JE (1952) External economies and diseconomies in a competitive situation. Econ J 62:54–67

Medema SG (1997) The trial of homo economicus: what law and economics tells us about the development of economic imperialism? In: Davis JB (ed) New economics and its history, suppl. 29. Duke university Press, Durham, pp 122–142

Medema SG, Zerbe RO Jr (2000) The Coase theorem. In: Bouckaert B, De Geest G (eds) Encyclopedia of law and economics, vol 1. Edward Elgar, Cheltenham, pp 836–892
Milgram S (1974) Obedience to authority. Harper & Row, New York
Peters HE (1986) Marriage and divorce: informational constraints and private contracting. Am Econ Rev 76(3):437–454
Plott CR (1983) Externalities and corrective policies in experimental markets. Econ J 93(369):106–127
Prudencio YC (1982) The voluntary approach to externality problems: an experimental test. J Environ Econ Manag 9(3):213–228
Rhoads TA, Shogren JF (1999) On Coasean bargaining with transaction costs. Appl Econ Lett 6(12): 779–783
Schwab S (1988) A Coasean experiment on contract presumptions. J Leg Stud 17(2):237–268
Shogren JF (1992) An experiment on Coasian bargaining over ex ante lotteries and ex post rewards. J Econ Behav Organ 17(1):153–169
Shogren JF, Moffett R, Margolis M (2002) Coasean bargaining with nonconvexities. Appl Econ Lett 9(15):971–977
Stigler GJ (1966) The theory of price, 3rd edn. Macmillan, New York
Szymanski S (2007) The champions league and the Coase theorem. Scott J Polit Econ 54(3):355–373
Vogel KR (1987) The Coase theorem and California animal trespass law. J Leg Stud 16(1):149–187

Codes of Conduct

George Tsourvakas
School of Journalism and Mass Communication, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract

Codes of conduct are self regulated rules adopted by people or organizations. All types of associations and organizations today have professional codes of conduct to avoid negative behaviors and to improve quality practices.

Keywords

Accountability; Best practices; Business ethics; Corporate social responsibility; Ethical code; Professional guide; Sustainable development

Synonyms

Accountability; Corporate social responsibility; Professional guides; Quality entrepreneurship; Related issues are business ethics; Sustainable development, and code of best practices (Schwartz 2004)

Definition

A code of conduct is both a sum of the ideal values or principals and a guide for good practice for individuals and organizations; it targets the reduction of negative externalities, the internalization of some social costs, and, at the same time, increasing benefits to society.

Theoretical Framework

Why do more and more organizations today try to follow, provide, and measure their social contribution based on a code of conduct? First, a code of conduct is a precaution mechanism. Following appropriate social behavior is less costly and risky than the costs of penalties, law suits, punishments, harm done to a brand name, customer boycotts, negative word of mouth, (Schwartz 2004).

Second, the strict enforcement of codes of conduct is a voluntary and self-regulated mechanism for companies that is not imposed by legislation and regulations. They are a signal and means to inform people that a company wishes to stay in the market for a long time and create a respected brand name; therefore, it is very important for companies to build a good image (Carroll and Buchholtz 2006).

Third, codes of conduct are an ex-ante strategy and not ex-post, which means that elements of positive ecological and social behavior could form part of a business strategy, giving companies the opportunity to think about how to operate in a way that offers something to society or without doing harm. This strategy could make companies very innovative and creative (Porter and Kramer 2006).

Fourth, companies that enforce codes of conduct in their policies and strategy significantly reduce their operational and transactional costs. This is because green companies save energy not only for society but themselves. Moreover, all types of stakeholders are happier to cooperate or work with socially responsible organizations, e.g., suppliers, distributers, shareholders, and bankers are more positive towards dealing with social responsible companies, reducing the cost of searching for, bargaining for, and enforcing contracts. Hence, employees who work in socially responsible organizations are more motivated and committed to the organizational dream and are professional in their work. Furthermore, codes of conduct can help organizations to achieve a balance between different stakeholder's interests (Thomsen 2001).

Fifth, companies that follow codes of conduct gain a positive reputation, and this is a guarantee to customers and citizens that they will not undertake opportunistic behavior such as higher prices, low quality, misleading practices, or intrusive behavior, but that they will respect them and show good will and trust. Thus, codes of conduct are a trust mechanism. Consequently, they make a positive contribution to the community, society, culture, and ecology development through payments, donations, and philanthropy for a better future. Therefore, socially responsible firms pursue a win–win strategy (Kotler and Lee 2005).

Empirical Evidence

Relevant cases that provide empirical evidence of this benefit are most national and international codes of conduct for hospitals and communication organizations or for individual doctors and journalists. According to their associations, there are behaviors that should be avoided, such as harassment and discrimination, advertising, disclosure of private information, vulgar language, being influenced by power, bribery, and corruption. At the same time, there are good practices that will benefit citizens and society, such as respecting people's dignity, explaining medical procedures, asking for permission for research, offering services voluntarily in emergency situations, and improving national culture and language.

Conclusion

All types of organizations could adopt altruistic corporate codes of conduct which complement legal requirements with a positive internal and external micro and macro environment.

Companies could care for all stakeholders, for instance employees, by providing training and healthy and safe conditions, and by discussing problems and respecting privacy, and at the same time offering sports, cultural, social, and ecological activities for their members. The same could be done for the micro external environment for customers, suppliers, distributors, and competitors. For instance, for customers, companies could offer safe products without their being produced by child labor or using animal testing, with logical prices, and with the products being eco-friendly and recyclable. Finally, with regards to the community, social, cultural, and physical macro environment, companies could incorporate into their professional codes proposals to solve social problems. For instance, communication organizations could support educational improvements and knowledge regarding educational problems that will change citizen's living conditions and lead to a better society.

References

Carroll AB, Buchholtz AK (2006) Business and society: ethics and stakeholder management, 6th edn. Thomson, Toronto

Kotler P, Lee N (2005) Corporate social responsibility: doing the most good for your company and your cause. Wiley, Hoboken

Porter M, Kramer M (2006) Strategy and society: the link between competitive advantage and social responsibility. Harv Bus Rev 84(12):78–92

Schwartz MS (2004) Effective corporate codes of ethics: perceptions of code users. J Bus Ethics 55(4): 323–343

Thomsen S (2001) Business ethics as corporate governance. Eur J Law Econ 11(2):153–164

Codes of Ethics

Luc Van Liedekerke[1] and Peter-Jan Engelen[2,3]
[1]Faculty of Applied Economics, University of Antwerp, Antwerpen, Belgium
[2]Utrecht School of Economics (USE), Utrecht University, Utrecht, The Netherlands
[3]Faculty of Business and Economics, University of Antwerp, Antwerpen, Belgium

Definition

A written set of guidelines issued by an organization to its stakeholders (primarily its employees) to help them conduct their actions in accordance with the primary values of the organization.

Introduction

The term "code of ethics" is often used interchangeably with other terms like code of conduct, deontic code, compliance code, integrity code, business code, etc. It is possible to range all these codes along a continuum with at the one end compliance-directed codes and at the other end ethics-directed codes. Compliance-directed codes are mainly focused on guaranteeing that its users stick to certain legal rules, while ethics codes start out from the values and norms of the organization. While this distinction is useful in order to characterize a code, almost all business codes today are a mixture of compliance and ethics. As will become clear in the historical section, there is also a clear legislative reason why most codes today are a combination of ethics and compliance.

Codes come in a huge diversity of form and content. An example that invited a lot of mockery is the 52 pages dress code launched by the Swiss bank UBS in 2010. It contained advice like "You can extend the life of your knee socks and stockings by keeping your toenails trimmed and filed" and discussed issues like eating garlic, the color of lingerie for women, and the length of thermic underwear (UBS 2010). On the other hand, we find examples of ethics codes that were crucial in the formation of an organization and continue to dominate its identity. A prime example is the Johnson and Johnson Credo. Launched in 1943, this one pager is still cherished by the company as the reference document that guides their decision-making (Johnson & Johnson 1943).

One could argue that the first ethics code was "do not eat the apple," and it immediately signals two fundamental issues with ethics codes: they are hard to implement and if broken imply considerable risk. Enron is a more recent example of the same. Its elaborate ethics code was launched in 2000 just one year before its spectacular collapse made abundantly clear that nobody had ever taken this code serious (Sims and Brinkmann 2003). Codes can be found in many places. Medical professions (Hippocratic Oath) and liberal professions are among the earliest to use codes of ethics as central element in self-regulatory efforts and still use them today (Higgs-Kleyn and Kapelianis 1999; Gaumnitz and Lere 2002). But not only barristers also pirates had their written codes of ethics (Kukla 2014). In this contribution we will not go into the many deontic codes for liberal and medical professions but instead concentrate on the history and use of ethics codes in business (Brooks 1989). We further narrow down the analysis to the Anglo-Saxon use of code of ethics and compliance as it can be argued that the Anglo-Saxon way of encoding ethics in business organizations has become the dominant form, certainly among large, stock-quoted companies (Rodriguez-Dominguez et al. 2009).

Evolution in Codes of Ethics

The history of ethics/compliance codes in (USA) business is driven by a cycle of scandal and regulatory response (Farrell et al. 2002). The first wave of scandals we would like to mention is best known through the Lockheed scandal but was not limited to Lockheed alone (Shaplen 1978). The Lockheed scandal encompassed a series of bribes and contributions made by Lockheed personnel in the process of negotiating the sale of aircraft. In fact in the mid-1970s, the US Securities and Exchange Commission

investigated over 400 US companies who admitted making questionable or illegal payments in excess of $300 million to foreign government officials, politicians, and political parties. In response President Carter enacted the Foreign Corrupt Practices Act (FCPA). This led to a first wave of compliance and ethics codes mainly among large, exporting companies with extensive contacts in the public sector. Defense industry scandals in the 1980s (defense contractors charged, for instance, 400$ for a simple hammer) lead to the Defense Industry Initiative (DII) that aimed and provided guidelines for building strong internal compliance programs in the defense industry (Kurland 1993). It is probably fair to say that this is one of the reasons why the defense industry has until today among the best-developed compliance and ethics programs.

A regulatory milestone for compliance programs was the reform of the Federal Sentencing Guidelines in 1991. These guidelines describe the elements of an organization's compliance and ethics program that are required to be considered for eligibility for a reduced sentence. It was the first time the government provided clear guidance on how a good compliance program needed to look like. After its introduction, it became clearly beneficial for companies to have a compliance program as judicial risk was directly linked to the presence of a decent compliance program (Ferrell et al. 1998). It resulted in a second, much larger wave of compliance programs in the USA, with expansions mainly in other Anglo-Saxon countries. Sentencing guidelines with reference to compliance were introduced in several countries among others the UK, France, Australia, Canada, Israel, and Singapore.

Around 2000 a new wave of scandals hit the USA with Enron undoubtedly the most familiar household name. During this time, Sarbanes-Oxley (SOX) regulation was the response by the regulator, and it contained, among other things, another review of the sentencing guidelines (published in 2004). The remarkable part about this revision is that every reference to the word "compliance" in the 1991 version was now replaced by "ethics and compliance." This fundamentally changed the target of compliance programs (Canary and Jennings 2008). It was no longer just "following the law;" instead creating an ethical culture inside the organization was now the goal which is vastly more complicated and uncertain. The reason for this change was that by 2000 many companies, most notably Enron itself, had established compliance programs, but the wave of scandals made painfully clear that these compliance programs failed to have any influence whatsoever on the organizational attitude toward the law (Sims and Brinkmann 2003). Clearly more and better compliance was needed. The change would have to reach deeper: the ethics of the organizations and its representatives needed to change.

The last wave of scandals was the financial crisis of 2008. Again governments reacted by regulatory change (Dodd-Frank in the USA, Mifid2 and Solvency2 in Europe, the Financial Instruments Exchange Act in Japan, etc.), with strong influence on compliance and ethics programs mainly in the financial sector.

In general one can say that over the past decades, ethics and compliance programs have increased dramatically. There are several reasons for this. First, there is a trend among legislators to push risks downward toward the individual firm. A good example from the financial sector is anti-money laundering (AML) regulation. While it used to be the regulator that had to look out for money laundering, now it is the financial service provider himself that carries the main responsibility to control for AML. Second, the cost of non-compliance has clearly increased over the past decades. Spectacular failures like Siemens, HSBC, or BNP each time involving billion dollar settlements have pushed companies to invest more in compliance and ethics programs. Third, the risk approach in business tends to take ethics and compliance much more serious and integrates it into an overall risk strategy.

As the regulatory environment becomes ever more complex, it is very likely that ethics and compliance departments will continue to grow. Nevertheless, ethics and compliance departments today remain effectively small. Most companies rely on one or at most a couple of full-time equivalents (FTE) to run their ethics departments.

Budgets are limited to at most a couple of million dollars, except for heavily regulated industries like finance, defense, and pharmaceutics where it can easily reach hundreds of millions.

Usefulness of Codes of Ethics

Despite the undeniable rise of ethics and compliance programs in business, it is hard to pinpoint any real progress with respect to integrity in business. For instance, survey after survey indicates that occupational fraud, corruption, and other forms of maleficence are not going down (Carberry et al. 2018). Ethics codes and compliance programs have therefore consistently been criticized from different sides. Managers resent the million dollar investment without clear return. Business ethicists have criticized the programs as at best shallow and at worst a hindrance to ethical conduct (Benson 1989; Frankel 1989; Stevens 1994; Painter-Morland 2010). A lot has to do with the crucial issue of how exactly a code of ethics is enacted inside the organization (Schwartz 2004; Munter 2013). Today implementation of ethics codes is primarily done through signing of the code (often part of the labor contract) and online training. Mandatory exercises on topics such as privacy, insider trading, and bribery are concluded by a ten-question quiz at the end. Many employees resent these mindless training sessions that are experienced as a series of box-checking routines, and according to business ethicists, instead of increasing ethics awareness, such an approach can actually invoke a cynical attitude toward ethics and refusal to take the code of ethics serious.

While this criticism is warranted, one should be realistic about the target of an ethics code (Lere and Gaumnitz 2003). If the target is a culture of integrity inside the organization (as is implied by the sentencing guidelines), it is impossible to reach such a goal only through compliance (Cressy and Moore 1983). It demands leadership at the top and efforts by audit, legal, HR, and several other departments that are connected to company culture (Sims and Brinkmann 2002). Compliance programs are just one element in the building of organizational integrity. Far more crucial, for instance, are incentive structures inside the organization. If commercial pressure is consistently high and reward is never linked to integrity indicators, one should not be surprised that compliance seems ineffective (Schwartz 2001). The danger today is that we end up with schizophrenic organizations with on the one hand increased control through ethics and compliance programs and on the other hand increased commercial pressure that drives ethical breaches. Such a schizophrenic organization simply breeds ethical cynicism toward the values and norms that the organization tries to push.

From this it follows that ethics and compliance programs should not be built nor judged in isolation (Collins 2012; Hoffman et al. 2001). They should be part of a general strategy that aims at corporate integrity (Chen and Soltes 2018). Today the success or failure of ethics and compliance programs is often measured in a very limited way. A survey by Deloitte in 2016 pointed out that the most common way is to measure completion rates of training programs and to deem training effective if enough employees – perhaps 90% or 95% – finish it (Deloitte 2017). Such an indicator mistakes legal accountability for compliance effectiveness. When the US Department of Justice published in 2017 an evaluation of corporate compliance in which this type of mistakes was pointed out, the document was immediately recuperated by business as a reference point that the justice department would from now on use to determine whether an ethics programs were effective (US DOJ 2017). It is another proof of the close interaction between regulation and the specific form of ethics and compliance programs.

Following a wave of accidents in the 1970s, the chemical sector launched the responsible care code and developed extensive health and safety programs. The impact of these programs has been clearly visible and measurable, for instance, in the number of workplace accidents. But this demanded a clear investment and consistent implementation with clear, measurable targets linked to financial incentives for the employee as well as for the company itself (e.g., the price of insurance was directly linked to accident rates). If codes of ethics want to have a clear impact on

corporate integrity, a similar road needs to be followed. Better measurement of ethics inside the organization, consistency with payment and promotion structures, pushing a speak-up culture in which employees are incentivized to speak about ethical issues, a demand to report ethical breaches under a no fault policy, follow up not just of things gone wrong but also of "near misses," and effective hotlines are all measures that companies can take in order to improve the impact of ethics codes. The benefits would not only reach the company but also the society in general.

Cross-References

- Audit Committees
- Codes of Conduct
- Corporate Criminal Liability
- Corruption
- Cost of Crime
- Organizational Liability
- Whistle-Blower Policy

References

Benson GC (1989) Codes of ethics. J Bus Ethics 8(5):305–319

Brooks LJ (1989) Corporate codes of ethics. J Bus Ethics 8(2–3):117–129

Canary HE, Jennings MM (2008) Principles and influence in codes of ethics: a centering resonance analysis comparing pre-and post-Sarbanes-Oxley codes of ethics. J Bus Ethics 80(2):263–278

Carberry EJ, Engelen PJ, Van Essen M (2018) Which firms get punished for unethical behavior? Explaining variation in stock market reactions to corporate misconduct. Bus Ethics Q 28(2):119–151

Chen H, Soltes E (2018) Why compliance programs fail – and how to fix them. Harv Bus Rev 2018:116–125

Collins D (2012) Business ethics: how to design and manage ethical organizations. Wiley, Hoboken

Cressey DR, Moore CA (1983) Managerial values and corporate codes of ethics. Calif Manag Rev 25(4):53–77

Deloitte (2017) In focus: 2016 Compliance Trends Survey. Online available at https://www2.deloitte.com/content/dam/Deloitte/us/Documents/governance-risk-compliance/us-advisory-compliance-week-survey.pdf. accessed on 14 June 2018.

Farrell BJ, Cobbin DM, Farrell HM (2002) Codes of ethics: their evolution, development and other controversies. J Manag Dev 21(2):152–163

Ferrell OC, LeClair DT, Ferrell L (1998) The federal sentencing guidelines for organizations: a framework for ethical compliance. J Bus Ethics 17(4):353–363

Frankel MS (1989) Professional codes: why, who and with what impact? J Bus Ethics 8(2–3):109–115

Gaumnitz BR, Lere JC (2002) Contents of codes of ethics of professional business organizations in the United States. J Bus Ethics 35(1):35–49

Higgs-Kleyn N, Kapelianis D (1999) The role of professional codes in regarding ethical conduct. J Bus Ethics 19(4):363–374

Hoffmann WM, Driscoll D, Painter-Morland M (2001) Integrating ethics. In: Moon C, Bonny C (eds) Business ethics: facing up to the interests. The Economist Book, London, pp 38–54

Johnson & Johnson (1943) The credo. Online available at https://www.jnj.com/credo. accessed on 14 June 2018.

Kukla R (2014) Living with pirates: common morality and embodied practice. Camb Q Healthc Ethics 23(1):75–85

Kurland NB (1993) The defense industry initiative: ethics, self-regulation, and accountability. J Bus Ethics 12(2):137–145

Lere JC, Gaumnitz BR (2003) The impact of codes of ethics on decision making: some insights from information economics. J Bus Ethics 48(4):365–379

Munter D (2013) Codes of ethics in the light of fairness and harm. Bus Ethics A Eur Rev 22(2):174–188

Painter-Morland M (2010) Questioning corporate codes of ethics. Bus Ethics A Eur Rev 19(3):265–279

Rodriguez-Dominguez L, Gallego-Alvarez I, Garcia-Sanchez IM (2009) Corporate governance and codes of ethics. J Bus Ethics 90(2):187

Schwartz M (2001) The nature of the relationship between corporate codes of ethics and behaviour. J Bus Ethics 32(3):247–262

Schwartz MS (2004) Effective corporate codes of ethics: perceptions of code users. J Bus Ethics 55(4):321–341

Shaplen R (1978) Annals of crime: the Lockheed incident. New Yorker 53:48–91

Sims RR, Brinkmann J (2002) Leaders as moral role models: the case of John Gutfreund at Salomon brothers. J Bus Ethics 35(4):327–339

Sims RR, Brinkmann J (2003) Enron ethics (or: culture matters more than codes). J Bus Ethics 45(3):243–256

Stevens B (1994) An analysis of corporate ethical code studies: "where do we go from here?". J Bus Ethics 13(1):63–69

UBS (2010) UBS Dress Guide for Client Advisors, Switzerland. Online available at https://static1.squarespace.com/static/55e0c62fe4b096b6619ff3d9/t/571bde1907eaa0d2558be4f6/1461444174809/Ubs+Dress+Code+English+2010.pdf. accessed on 14 June 2018.

US DOJ (2017) Evaluation of corporate compliance programs, US Department of Justice, Washington, DC. Online available at https://www.justice.gov/criminal-fraud/page/file/937501/download. accessed on 14 June 2018.

Cognitive Law and Economics

Angela Ambrosino[1] and Marco Novarese[2]
[1]Department of Economics and Statistics Cognetti de Martiis, University of Turin, Torino, Italy
[2]Department of Law and Economics, University of Eastern Piedmont, Centre for Cognitive Economics, Alessandria, Italy

Definition

The tendency to consider the Behavioral Law and Economics and Cognitive Law and Economics as different sides of the same coin has been widespread inside the discipline. That was the consequence of a miscomprehension of what behavioral economics and cognitive economics are. These two research areas arise from a shared critique to standard neoclassical economics assumption of agents' perfect rationality and a common idea that economic agents, in the real world, are heterogeneous and more cognitive complex than what the theory assumed, but soon they diverge pursuing different goals and partially applying different research tools. Particularly BL&E is more concerned with what agents do, while CL&E is more about how agents think.

Hence we need a proper discussion of what Cognitive Law and Economics is as well as we need a proper definition of Behavioral Law and Economics.

Introduction

Do we really need an autonomous definition for Cognitive Law and Economics or it is the same of Behavioral Law and Economics? The tendency to consider the two approaches as different sides of the same coin has been widespread inside the discipline. That was the consequence of a miscomprehension of what behavioral economics and cognitive economics are. These two research areas arise from a shared critique to standard neoclassical economics assumption of agents' perfect rationality and a common idea that economic agents, in the real world, are heterogeneous and more cognitive complex than what the theory assumed, but soon they diverge pursuing different goals and partially applying different research tools. Hence we need a proper discussion of what Cognitive Law and Economics is as well as we need a proper definition of Behavioral Law and Economics.

Other entries in this encyclopedia show how and when law meets economics (see Law and Economics or Behavioral Law and Economics or Nudge or Financial Education). When law scholars started applying the insights offered by neoclassical economics to their inquiry, the aim of this new approach to law was to develop both a positive and a normative theory of law on which to build efficient legal norms. Law and economics (L&E) uses economic models and econometric tools to develop its research in two ways:

1. Pursuing efficiency: efficiency is considered from two different points of view; on the one hand, it means that common law (judge-made law) is efficient, and on the other, from a normative point of view, it also means that law must be efficient.
2. Its emphasis on incentives and people's responses to those incentives.

L&E has been widely criticized (Ellickson 1989) in that applying economic tools is not sufficient to investigate the logic underlying the law and that the reductionist approach of economics cannot enable L&E to develop a proper positive theory of law and it excludes any consideration about justice.

L&E has been strongly influenced by the changes and debates that have characterized the development of economics since the middle of the last century (Rachlinski 2000). In recent years, the results obtained by the behavioral economics have given new emphasis to the first criticisms brought against law and economics. Behavioral economics shows that human behavior deviates from the perfect rationality assumption, and these deviations are not completely random, so it is possible to model and predict human behavior when it is affected by biases. During the 1990s, Jolls et al. (1998) investigate the opportunities offered by

behavioral economics to develop a new approach to law based on a more exhaustive theory of human behavior whereby better understanding of the foundations of individual behavior should strengthen both the descriptive power of models and their normative power. Their pioneering work gives rise to Behavioral Law and Economics (BL&E). During these same years, inside economics is developing another important research approach called cognitive economics (CI) (Bourgine and Nadal 2004). Cognitive economics shares with the behavioral approach the idea that human behavior is complex and that economic theory must ground its theories on a better understanding of cognitive decision-making processes. Cognitive economics retrieve the tradition of what Sent (2004) define "Old Behavioral Economics" that is the approach by Herbert Simon, instead that Kahneman's.

Nevertheless, the two approaches follow (almost partially) different paths of inquiry. Cognitive economics puts itself in opposition to neoclassical economics investigating economic problems as complex phenomena. Its inquiry focuses on the analysis of the micro-foundations of human behavior and applies an interdisciplinary approach. Cognitive economics strongly criticizes the assumptions of standard economics and focus on the complexity of decision-making processes of heterogeneous agents. It questions the predictions of standard economics models and the rigidity of the formal tools applied. It is aimed at understanding decision-making processes, but it differs from behavioral economics, whose methodology is based on the analysis of the effectively exhibited behaviors. Cognitive economics' central idea is that each phenomenon can be investigated with different tools and from different points of view. For example, cognitive economics investigates interdependent decisions using game theory not as a formal tool to predict specific outcomes but as a framework of analysis that allows investigating the complexity of agents' decision-making processes (Schelling 1960); the outcomes of the game do not simply depend on strategies, but they are strongly linked to social context, path dependence dynamics, and focal pints. Cognitive economics focus on norms and institutions (Rizzello and Turvani 2000, 2002), but while law and economics has been much influenced by behavioral economics, the cognitive analysis of institutions has not been considered until recently.

Ambrosino (2016) shows two main explanations for this lack of interest in the cognitive theory of institutions:

1. The different concept of norms underlying the two research fields.
2. The cognitive theory of institutions is still far from developing a normative theory, and it focuses its inquiry on the positive level.

Nevertheless in the last few years, part of the literature points out the relevance of the analysis of the role of institutional forces and social norms in constraining and coordinating heterogeneous individuals, and cognitive economics and law and economics start to be connected and a new path of inquiry is arising.

The next sections are organized as follow: section "Why Behavioral Law and Economics is not Cognitive Law and Economics" explains why Cognitive Law and Economics (CL&E) is not the same as BL&E, particularly, "Toward a Cognitive Approach to Law and Economics" describes the main feature of CL&E, and "Main Critiques to Behavioral Law and Economics" focuses on the main critiques that such approach moves to behavioral law and economics. Section "Toward a Cognitive Law and Economics Inquiry" provides an example of how CL&E contributes to the inquiry into law.

Why Behavioral Law and Economics is not Cognitive Law and Economics

Toward a Cognitive Approach to Law and Economics

The cognitive theory of institutions is grounded on the idea that it is not possible to investigate the rise and evolution of institutions without investigating individual decision-making processes (North 2005). The institutional and the individual levels of analysis are interconnected, so that an institutional change may be the starting point for

modification of agents' behavior, and new cognitive classifications or new routines of behavior can engender a slow process of institutional change (Hayek 1982; Hodgson 2004; Ambrosino 2014). Cognitive theory of institutions assumes that agents are heterogeneous. Heterogeneity means that agents can exhibit different behaviors even if they belong to the same social and cultural context. That heterogeneity doesn't prevent coordination because agents are different, but they are made up of the same ingredients (Hayek 1982). Hence, they are able to understand each other, to build correct expectations about each other's behavior, and to share common social norms.

Recently such research filed shows points of contact with that part of the legal theory that firmly critiques BL&E. Such connection opens the door to a proper cognitive approach to L&E.

Particularly, Gregory Mitchell's main works seems to represent the main contribution to developing inquiry into the "individual-institution" framework already described by the cognitive theory of institutions (Hodgson 2004; Ambrosino 2014). Mitchell's critique of BL&E "provides reasons why legal theory should refrain from broad statements about the manner in which all legal actors process information, make judgments and reach decisions and why others should be skeptical of such broad claims by the legal decision theorists" (2002b, p. 33); "legal decision theorists should recognize the need for greater caution and precision in drawing of descriptive and prescriptive conclusions from empirical research on judgment and decision making" (2002b p. 32). Mitchell's contribution is based on a strong belief in the utility of psychological and other empirical research for legal analysis.

It emerges a new approach to law that shares with cognitive institutional economics the idea that agents are heterogeneous and that simply introducing the existence of "standard" biases in modeling human behavior does not enable the development of efficient predictive models; the perfect rationality assumption is not an appropriate instrument with which to investigate agents' behavior, and a proper theory of human behavior is needed. This approach suggests that the existence of cognitive biases in legal contexts must be investigated in the field and with respect to specific contexts through "social facts studies" (Mitchell et al. 2011): a social facts study applies different research methods to explain case-specific descriptive or causal claims, and it is focused on the context-specific features of the case at hand. The analysis of how agents should behave cannot be separated from the investigation of the specific social context and cultural and social relations. A multidisciplinary approach is necessary to develop better inquiry into the complexity of decision-making processes in legal contexts. Legal theory, hence, moves toward a new approach, in which the cognitive determinants of agents' behavior are investigated; it highlights the importance of (i) agents' cognitive predispositions, (ii) learning processes and the influence of past experience, and (iii) the role of context. Moreover a cognitive inquiry into the diffusion of normative behavior and institutional change can furnish key into the opportunities offered by the development of prescriptive rules in shaping individual behavior. It emerges a new metacognitive approach to legal theory in which norms are concrete instruments with which to induce agents to develop different ways of processing information.

CL&E, following a social facts analysis, shows how to build appropriate decision tools based on objective casual claims. Scientific research results can be applied to normative purposes. They should constitute a sort of "social authority": an organizing principle for courts' of legislator' use of social science to create or modify a rule of law (Monahan et al. 2009). In the perspective of CL&E, social research and legal theory partially lose the need to furnish normative models. Producing case-specific evidence through reliable social science principles and methods, they become the research instruments that give judges and courts, and more generally the legislator, the information and the tools with which to evaluate and create new rules of law.

Main Critiques to Behavioral Law and Economics

Part of the literature inside legal theory criticizes BL&E both under a theoretical and a methodological point of view and points out relevant

elements of contact with cognitive economics that has opened the door to a new path of inquiry.

BL&E arise to pursue two main aims: first, explain why people do not act as they should in context of interest for legal theory (the benchmark being that agents should behave as the perfect rationality assumption expects), and second, bring people to act as they should proposing "a form of paternalism, libertarian in spirit, that should be acceptable to those who are firmly committed to freedom of choice on grounds of either autonomy or welfare" (Sunstein and Thaler 2003, p. 1160).

To pursue such aims, BL&E applies the tools and the insights furnished by behavioral economics. It is not surprising that BL&E today is exposed to quite the same critiques as behavioral economics (Ambrosino 2016).

The first critique to BL&E is strictly related to one of the cornerstone ideas inside B&E. It is a common opinion in B&E that it is possible to incorporate the complexity of the cognitive determinants of human behavior into the standard formal models of the neoclassical approach. The idea is that the assumption of perfect rationality can be replaced with a new concept of rationality – in which the existence of deviations from the perfect rationality assumption is explained by introducing new variables corresponding to particular biases assumed as commonly shared among agents – that better explains the complexity of real decision-making processes. Behavioral economics returns to being a research approach completely compatible with mainstream economics (Davis 2013). This tendency to build formal models has also taken place in the behavioral approach to L&E (Korobkin and Ulen 2000). The replacement of the perfect rationality assumption guarantees that BL&E models, compatible with the mainstream, produce strong normative outcomes. The first criticism to BL&E concerns the way in which scholars introduce into their inquiries insights drawn from the cognitive and psychosocial research of the past 30 years (Mitchell 2002a, 2002b, 2003a). BL&E grounds its research on the evidence of the existence of cognitive biases in human behavior and argues that such biases are widespread in the population and are responsible for predictable and systematic errors (Korobkin and Ulen 2000). Nevertheless BL&E scholars fail in their attempt to criticize the perfect rationality assumption because they do not develop a new concept of rationality including the complexity of human decision-making processes. BL&E substitutes the neoclassical assumption of perfect rationality with an assumption of "equal incompetence" (Mitchell 2002a). This assumption is based on empirical research that shows homogeneous behavioral tendencies among agents. BL&E uses these behavioral tendencies to compile a list of common deviations from rationality that characterizes the entire population, and it develops normative models prescribing how agents have to behave and how decision-makers should intervene to shape agents' behavior and avoid their errors. B&LE overlooks the substantial empirical evidence that people are not equally irrational and that human behavior is strongly influenced by situational variables: "The only way the lessons of behavioral decision research on bounded rationality can be manageably incorporated into behavioral models for use in the law is if these lessons apply widely and uniformly. If the rationality of behavior depends on particular characteristics of the legal actor or on even just a few characteristics of the situation at hand, then the development of behavioral models that are both realistic and predictive becomes enormously complex" (Mitchell 2002a p. 83). CL&E argues that BL&E do not understand that heuristic processing is only one mode of thought and that agents often do not act as expected, and it suggests the need for a legal theory focused on finding solutions to specific problems rather than on developing a general model of legal behavior. Heuristics can lead to favorable solutions but in many cases they can also give rise to errors. BL&E relies on the results obtained by behavioral research developed in other branches of economic theory and generalizes their significance. One of the main contributions is the pioneering work of Kahneman and Tvresky (1974). These authors argue that their "studies on inductive reasoning have focused on systematic errors because they are diagnostic of

the heuristics that generally govern judgment and inference" (1974, p. 313). But this does not mean that the so-called K-T man can be reduced simply to the use of rules of thumb and heuristics in judgment. It seems an excessively simple explanation of human decision-making. "The likelihood that a particular decision or judgment will deviate from the ideal behavior derived from norms of rationality depends on a range of personal and situational factor. Even inside the relatively controlled environment of the laboratory, we see considerable variation in cognitive performance among individuals depending on their cognitive abilities, educational background, and affective state" (Mitchell 2002a, p. 109). CL&E suggests legal theory should not seek a general model of judgment and decision-making, but it should develop a contextualist approach that seeks to identify the conditions under which irrational behavior occurs. BL&E has important normative, methodological, and empirical limitations that prevent it from achieving descriptive and predictive accuracy. The libertarian paternalism suggesting that planners can improve social welfare by setting default rules that create benefits for those who commit errors but cause little or no harm to those who are fully rational (Sunstein and Thaler 2003) assumes the pervasiveness of irrational tendencies but ignores less invasive forms of intervention that may help agents overcome their errors without altering the substantive rights of the parties (Mitchell 2005). BL&E describing behavior as rational or irrational requires a normative standard against which the behavior may be judged (Mitchell 2003b). The behavioral approach assumes that rationality requires logical consistency and coherence in the formation and ordering of beliefs and preferences (Kahneman 1994; Simon 1997). Rationality as coherence operates as a closed system. Individual defines goals and beliefs and behavior must be logically consistent and coherent with respect to those goals and beliefs. In the case of legal judgment, when evidence of an irrational judgment is found, many different explanations are possible, some of which make the irrationality of the decision questionable (Mitchell 2003b). A behavior in a particular context may be at the same time rational and irrational depending on the goals, the interpretation of the situation, and the tools used by any agent involved in the decision-making process.

The second main criticism concerns the methods employed to test for cognitive biases and errors (Mitchell 2002b, 2003b). BL&E research underestimates situational and individual variations in behavior and employs relatively weak tests of the hard-core assumptions of agents' cognitive feature. The point is that the core of the research in heuristics and biases is based on statistical significance tests on experimentally generated and aggregate data. This body of research formulates in general terms the conditions under which events of various sorts occur and provides an interesting set of findings in general terms but with unspecified practical implications. Judgments are summarized by averaging across all the experimental subjects. That means that in BL&E analysis, if individual differences among judges emerge, these differences are treated as "errors," and an "average judge" is considered the most meaningful summary of judges. This approach has the advantage of ensuring generalizability. Therefore, rather than examining individual variation in judgment and choice, behavioral decision theorists typically assume that "to a first approximation, the thought processes of most uninstitutionalized adults are quite similar, and any variation in subjects' responses is attributed to measurement error or random variance" (Mitchell 2002b, p. 46). The rigor of experimental research is purchased at the price of generalizability of results, and this trade-off operates most directly in those fields that use laboratory experiments to study how humans navigate complex social environments like BL&E. Such critique is strongly related to the debate emerged in psychology about the danger of relying on "statistical significance" as a measure of behavioral tendencies. Scientists (and journals) publish studies and analyses that "work" and place those that do not in the file drawer (Rosenthal 1979). One answer to this problem of publication bias is that we can trust a result if it is supported by many different studies. But this argument breaks down if scientists exploit ambiguity in order to obtain statistically significant results (Simmons et al. 2011).

Toward a Cognitive Law and Economics Inquiry

Hence Cognitive Law and Economics is aimed at developing a legal theory in which the peculiarity of decision-making in legal contexts can be really explained. The critique of the equal incompetence assumption suggests the need for a new analysis in which heterogeneous agents are considered (Mitchell 2002a, b, 2003a, b).

Evidence on cognitive biases must be investigated in legal contexts so as to build an original and consistent map of evidence. CL&E aspires to develop a contextualist approach. A contextualized approach acknowledges that features of the person, the situation, and the task have an impact on the nature and quality of judgment.

Experiments are only one of the tools that should be applied to examine variations in individual behavior. The need for an interdisciplinary approach arises from the recognition that multiple forces combine to produce particular behaviors. The cognitive theory of institutions has yet developed interesting inquiries into coordination processes (Schelling 1960) and into the relevance of learning in the process through which people conform to social or formal rules.

More recently, an example of the kind of inquiry CL&E can develop is given by Mitchell (2009) idea of a metacognitive approach to regulation. Such approach is based on his discussion about the role of second-level thought in shaping human behavior. BL&E describes judgment as the product of a non-deliberative thought process based on cognitive heuristics and rules of thumb. Psychological models of actors developed inside BL&E show that biases in judgment and errors often arise at the level of first-order thoughts; thoughts occur at the direct level of cognition and are not intentional and not deliberative. These models assume that agents are incapable of going beyond these first-order thoughts and that this is the cause of irrational and discriminatory behavior. This emphasizes the role of automatic and intuitive thoughts while neglecting the role played by controlled and deliberative thoughts. It leaves no room for self-correction, arguing that individuals lack self-awareness of their biases, and it ignores the substantial evidence that agents learn through experience. Second thoughts may be the products of conscious effort, but they may also be automatic corrections working at the unconscious level. The propensity to engage in self-correction varies among persons and situations, but all cognitively normal people are able to engage in some amount of "metacognition" about their own thoughts (Loires 1998). People may differ in their propensity for such reflection depending on their education, upbringing, values, or genetic endowment, but everyone possesses some level of ability in rethinking their own thoughts.

Regulation should take it into consideration. If second thoughts apply, law will not simply change the prices of different behaviors for the purposes of a rational analysis of the costs and benefits of different courses of action. Rather, law will focus on altering the ways in which agent processes information. Under this point of view, law is a system of second thoughts that functions both consciously and unconsciously. Hence, law can contribute to influencing thoughts and behaviors in legal contexts. Mitchell provides concrete applications of his theory of law. The author (Monahan et al. 2009; Mitchell 2010; Mitchell et al. 2011) enters the debate on the proper scope of expert witness testimony that purports to summarize general social science evidence to provide context for the fact-finder to decide case-specific questions. Mitchell's analysis focuses on the Dukes v. Wal-Mart case on gender discrimination toward female employees. Dukes' plaintiffs submitted expert statistical evidence showing that female employees were faring worse in the aggregate than male employees, and a report by a social science expert identified a common source of this discrimination across all Wal-Mart facilities (Mitchell 2010, p. 136). The social science expert based his report on the "social framework analysis" method (Fiske and Borgida 1999). This method consists in using social science research as a framework for analyzing the facts of a particular case. The reliability of such analysis is based on the reliability of the research on which the general conclusions applied to the case at hand are based. In Dukes v. Wal-Mart, the expert

summarized research on gender bias, organizational culture, and anti-discrimination measures and applied it to interpret the facts in the discovery material supporting the claims of the Dukes plaintiffs. Mitchell argues that testimony based on that social framework analysis should be restrained from making any linkage between general social science research findings and specific case questions. In the specific case of Dukes v. Wal-Mart, he based his critique on two main points: (1) in social framework analysis, experts use their personal judgment rather than scientific method to link social science to specific cases; in some sense, social framework analysis make the same mistake that BL&E does in extending the experimental economics results to its research purposes without dealing with context-specific research. (2) The expert corroborated his report with statistical evidence. But the statistical evidence was itself subject to dispute with regard to the proper unit of analysis. The plaintiffs argued for an aggregate-data approach. This choice did not allow consideration of context-specific differences due to store-by-store variation in male-female outcomes and to local control over personnel matters. This use of statistical evidence is an example of how statistical results can vary depending on the many decisions that researchers have to make while collecting and analyzing data (which outliers to exclude, which measures to analyze, and so on). Mitchell argues that there are social science techniques and methods that allow development of opinions about the parties or behaviors involved in a particular case; such evidence has been referred to as "social facts" (Mitchell et al. 2011). Social facts are special types of adjudicative facts produced by applying social science techniques to case-specific data in order to help prove some issue in the case. A wide variety of social science methods can be used to produce social facts. The design of a social fact study depends on what a party hopes to learn. Mitchell divides the search for social facts according to three main goals:

1. Obtaining descriptive information: getting the facts right is important, but doing so can be difficult when the relevant facts are in the possession of a large number of nonparties.
2. Obtaining explanatory information: gain a better understanding of the issue in a case. Many research methods can be applied, such as interview, survey, observational study, and experimental simulation.
3. Testing specific hypotheses: the ideal way to test causal hypotheses is through the use of experiments in which participants' behaviors are recorded to assess how changes in the experimental conditions affect the behavior in question (Mitchell et al. 2011).

Social facts constructed by a proper scientific method possess scientific reliability and fit the facts of a particular case. Such reliability depends on the reliability of the scientific method applied. Mitchell shows that when addressing such a complex task as deciding a legal dispute, it is necessary to rely on rigorous interdisciplinary research tools that help prove some issue in the case.

CLE remains a very recent research project; its finding can be still considered a preliminary attempt to develop a proper interdisciplinary inquiry to law and economics. Moreover this approach is still mainly focused on a positive ground. As shown in this section, CL&E is a very relevant and promising research field.

Cross-References

- ▶ Behavioral law and economics
- ▶ Law and Economics
- ▶ Nudge

References

Ambrosino A (2014) A cognitive approach to law and economics: Hayek's legacy. J Econ Issues 48(1):19–49

Ambrosino A (2016) Heterogeneity and law: toward a cognitive legal theory. J Inst Econ 12(2):417–442

Bourgine P, Nadal JP (eds) (2004) Cognitive economics: an interdisciplinary approach. Springer, London

Davis JB (2013) Economics imperialism under the impact of psychology: the case of behavioral development economics. Oeconomia 1:119–138

Ellickson RC (1989) Bringing culture and human frailty to rational actors: a critique of classical law and economics. Chicago Kent Law Rev 65:23–55

Fiske ST, Borgida E (1999) Social framework analysis as expert testimony in sexual harassment suits. In Estreicher S (ed.) Sexual harassment in the workplace: proceedings of New York University 51st annual conference on labor. New York, pp 575–577
Hayek FA (1982) Law, legislation and liberty. Routledge, London
Hodgson GM (2004) Reclaiming habit for institutional economics. J Econ Psychol 25:651–660
Jolls C, Sunstein CR, Thaler R (1998) A behavioral approach to law and economics. Stanford Law Rev 50:1471–1552
Kahneman D (1994) New challenges to the rationality assumption. J Inst Theor Econ 150:18–36
Kahneman D, Tvresky A (1974) Judgment under uncertainty: heuristics and bias. Science 185:1124–1131
Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. Calif Law Rev 88(4):1051–1144
Loires G (1998) From social cognition to metacognition. In: Yzerbyt VY, Loires G, Dardenne B (eds) Metacognition: cognitive and social dimensions. Sage, London, pp 1–15
Mitchell G (2002a) Why law and economics' perfect rationality should not be traded for behavioral law and economics' equal incompetence. Georgetown Law J 91:67–167
Mitchell G (2002b) Thinking behavioralism too seriously? The unwarranted pessimism of the new behavioral analysis of law. William Mery Law Rev 43:1907–2021
Mitchell G (2003a) Tendencies versus boundaries: levels of generality in behavioral law and economics. Vanderbilt Law Rev 56:1781–1812
Mitchell G (2003b) Mapping evidence law. Michigan State Law Review, 1065–1148
Mitchell G (2005) Libertarian paternalism is an Oxymoron. Northwest Univ Law Rev 99(3):1245–1277
Mitchell G (2009) Second thoughts. McGeorge Law Rev 40:687–722
Mitchell G (2010) Good causes and bad science. Vanderbilt Law Rev Banc Roundtable 63:133–147
Mitchell G, Monahan L, Walker L (2011) Case-specific sociological inference: meta-norms for expert opinions. Sociol Methods Res 40:668–680
Monahan J, Walker L, Mitchell G (2009) The limits of social framework evidence. Law Probab Risk 8(4): 307–321
North D (2005) Understanding the process of economic change. Princeton University Press, Princeton
Rachlinski JJ (2000) The "New" law and psychology: a reply to critics, skeptics, and cautious supporters. Cornell Rev 85:739–766
Rizzello S, Turvani M (2000) Institution meet mind: the way out of an impasse. Constit Polit Econ 11:165–180
Rizzello S, Turvani M (2002) Subjective diversity and social learning: a cognitive perspective for understanding institutional behavior. Constit Polit Econ 13:201–214
Rosenthal R (1979) The file drawer problem and tolerance for null results. Psychol Bull 86:638–641
Schelling TC (1960) The strategy of conflict. Harvard University Press, Cambridge, MA
Sent EM (2004) Behavioral economics: how psychology made its (limited) way back into economics. Hist Polit Econ 36:735–760er
Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22:1359–1366
Simon H (1997) Models of bounded rationality. MIT Press, Cambridge, MA
Sunstein C, Thaler R (2003) Libertarian paternalism is not an Oxymoron. Univ Chicago Law Rev 70:1159–1202

Coherence

▶ Rationality

Collective Redress

▶ Class Action and Aggregate Litigations

Commercial Benefits

▶ Preferential Tariffs

Commercial Concessions

▶ Preferential Tariffs

Commercial Sex and Health

Giovanni Immordino and Francesco F. Russo
Department of Economics and Statistics, University of Naples Federico II and CSEF, Naples, Italy

Abstract

The spread of HIV/AIDS and of other sexually transmittable infections is affected by the legal regime governing prostitution. In this entry we briefly discuss a theory of commercial sex, public intervention, and the health consequences. We present some evidence and the

main motivations for unprotected commercial sex in addition to examining the theoretical and empirical effects of different policies for the health consequences of commercial sex.

Definition

Link between the legal regime governing commercial sex and health.

Introduction

The spread of HIV/AIDS and of other sexually transmittable infections (STI henceforth) is an important negative externality of commercial sex. The feasibility and effectiveness of public intervention differ substantially depending upon the legal regime in place. The reason for this is that different policies affect the endogenous evolution of the risks in the market for prostitution services differently. Any policy intervention will change the identity of the clients and of the sex workers who decide to join, or not join, the market, and it will also have an impact on the decision regarding which market to join (legal or illegal). This will endogenously change the risks associated with the markets, thus further influencing demand and supply, and consequently the identities, and so on. This is the reason why even well-meant interventions can lead to unintended consequences which not only reduce the effectiveness of the intervention but sometimes even make things worse.

Finally, we should be aware that legal regime governing prostitution might have other consequences besides affecting public health. For instance, Immordino and Russo (2015b) show that if prostitution is legal or regulated, individuals tend to justify it significantly more than if it is prohibited.

A General Theory of Commercial Sex and Health Consequences

Immordino and Russo (2015a) look at a specific negative externality associated with prostitution, the spread of HIV/AIDS and of other STIs, from a general and theoretical point of view. In particular, they set up an equilibrium model of prostitution and calibrate it to real-world data to perform a quantitative policy analysis exercise. They model demand and supply of prostitution, given the policies implemented by the government, as a function of the risk of infection, of the health status, and of the outside earning opportunities for both clients and sex workers. They consider the policies within three different regimes. In a laissez-faire regime, prostitution is legal and unlicensed and no policy is implemented. In a *prohibition* regime, prostitution is illegal and the enforcement consists of random audits and fines. Under *regulation*, prostitution is legal, conditional on compliance with a series of requirements such as tax payments, entry fees, or participation restrictions. Two markets coexist in this regime: a legal market, where all the agents comply with the regulations, and an illegal market where they do not. The enforcement against the illegal sector consists of random audits and fines. Key features of this model are the interaction between the demand and supply of legal and illegal prostitution and the endogenous evolution of the risks: any policy intervention influences the characteristics of clients and sex workers that join the two markets and, therefore, the risks associated with buying and supplying prostitution, further influencing demand and supply and so on. For instance, introducing mandatory health checks for sex workers decreases the probability of infection for customers, thereby attracting more demand for prostitution by healthy customers. This in turn decreases the average risk of infection for sex workers, thereby inducing more healthy individuals to join the supply side of the prostitution market and so on.

Motivations for Unprotected Commercial Sex

Unprotected sex increases the risk of HIV/AIDS and the transmission of other STIs. It is therefore extremely important to understand why sex

workers and clients might engage in unprotected sex. Concerning sex workers, the main reasons are lack of awareness about HIV/AIDS and about safe sex practices (Bhave et al. 1995), the high cost and poor availability of condoms, violent practices by clients who force them to have unprotected sex, and extra compensation that some customers are willing to pay for unprotected sex. With respect to clients, the main reasons are intrinsic preference not to use condoms and a lack of information about the risks.

Bhave et al. (1995), in a study of Bombay sex workers, show that monetary incentives are the most important factor deterring sex workers from using condoms, even for those who are aware of the risks of unprotected sex and of the preventive role of condoms. The importance of monetary incentives is also documented in two studies that use micro data to quantify the risk compensation required by sex workers for not using condoms. The conclusion of both studies is that sex workers accept the extra risk because clients are willing to pay them very well for it (Rao et al. 2003; Gertler et al. 2005). Specifically, Rao et al. (2003) studied a random sample of commercial sex workers from the red-light area of Sonagachi in Calcutta. Using price data for transactions with and without condoms, they were able to estimate the compensating differential for safe sex. They found a loss in average prices between 66% and 79% in the case of condom use. Similarly Gertler et al. (2005), using data from Mexico, estimate a 23% premium for unprotected sex (46% in case of a "very attractive" sex worker). Overall, the strong monetary disincentive for safe sex practices can have a substantial adverse impact on preventing the spread of HIV/AIDS and of other STIs.

Moving to the demand side of the market, Cameron and Collins (2003) use UK micro data to identify the male clients' characteristics that influence their demand for sex services. They find, among other things, that a high risk of HIV/AIDS infection has a negative impact on sex service usage. Their results show support for the view of "the man who pays for sex as a rational economic actor," suggesting that interventions on the side of the client might be effective.

Effects of Public Interventions: Theory and Evidence

The conventional policy recommendations to reduce unprotected sex are to intervene on the supply side of the market by implementing educational campaigns, improving access to inexpensive condoms, enforcing laws against human trafficking and rape, etc. Moreover, since an important motivation for unprotected commercial sex is the clients' willingness to pay, complementary interventions on the demand side are also necessary in order to increase condom use.

However, the legal regime in place matters a great deal with respect to the effectiveness of different policies on the health consequences of commercial sex. Immordino and Russo (2015a) show theoretically that in a *prohibition* regime, targeting enforcement at customers only has a worse outcome on reducing the negative health externality than targeting enforcement at sex workers only. They also show that, in a *regulation* regime, taxation can increase harm. The reason is that higher taxes, by raising the cost of operating legally, decrease the number of healthy individuals that join the legal market, which is, therefore, smaller but riskier. Conversely, increased enforcement against the illegal market increases the equilibrium quantity in the legal market, also through a decrease in risk. Establishing a licensing system that prevents risky individuals from joining the legal market, on the other hand, reduces the risk and, therefore, increases the demand for legal prostitution. The impact on harm is theoretically ambiguous, but, for many model parameterizations, the decreased risk prevails so that harm also decreases. Finally, in comparing regulation and prohibition, they find that regulation is better for harm minimization, while the laissez-faire regime is always dominated (see also Immordino and Russo 2016).

Gertler and Shah (2011) studied empirically the effects of different policies in Ecuador, where a *regulation* regime prevails. Prostitution is legal, but sex workers must obtain and periodically renew a license at their own expense. This license, among other requirements, is conditional on health status and, in particular, on being free

from STIs. Licensed prostitutes can solicit in nightclubs and bars or in the streets, and typically, licensed and unlicensed prostitutes share the workplace in both nightclubs and in the streets. The authors exploit a nationwide dataset that combines sociodemographic, economic, and health information for the sex workers and regional, within-country variation of police enforcement against unlicensed prostitutes in the streets and in bars. They find that increased enforcement against unlicensed street prostitution is associated with a smaller diffusion of STIs among sex workers, which is in line with the theoretical results in Immordino and Russo (2015a). Gertler and Shah (2011) also show that public intervention can lead to unintended consequences. Indeed, enforcement against *unlicensed street prostitution* pushes some sex workers to move to the bars where there is a lower demand for unprotected sex and lower STI rates among clients, with the result that the overall risk of infection decreases. However, increased enforcement against *unlicensed prostitution in bars and nightclubs* is associated with higher infection rates. The reason is that an increase in enforcement raises the cost of being a sex worker without a license in a bar with respect to both complying with the license and with being unlicensed on the streets. Therefore, the question is if, in response to a tighter enforcement, the unlicensed sex workers will choose to obtain and renew the license or to move to the streets where unprotected sex is more common and STIs among clients more frequent. Gertler and Shah (2011) show that migration toward the street sector prevails, with an overall increase in the infection rate.

Cross-References

- ▶ Crime and Punishment (Becker 1968)
- ▶ Liberty
- ▶ Pornography
- ▶ Prostitution, Demand and Supply of
- ▶ Sex
- ▶ Sex Offenses
- ▶ Shadow Economy
- ▶ Underground Economy

References

Bhave G, Lindan CP, Hudes ES, Desai S, Wagle U, Tripathi SP, Mandel JS (1995) Impact of and intervention on HIV, sexually transmitted diseases, and condom use among sex workers in Bombay, India. AIDS 9(Suppl 1):S21–S30

Cameron S, Collins A (2003) Estimates of a model of male participation in the market for female heterosexual prostitution services. Eur J Law Econ 16: 271–288

Gertler P, Shah M (2011) Sex work and infection: what's law enforcement got to do with it? J Law Econ 54(4): 811–840

Gertler P, Shah M, Bertozzi SM (2005) Risky business: the market for unprotected prostitution. J Polit Econ 113: 518–550

Immordino G, Russo FF (2015a) Regulating prostitution: a health risk approach. J Public Econ 121:14–13

Immordino G, Russo FF (2015b) Laws and stigma: the case of prostitution. Eur J Law Econ 40: 209–223

Immordino G, Russo FF (2016) Optimal prostitution policy. In: Cunningham S, Shah M (eds) Handbook of the economics of prostitution. Oxford University Press, Oxford, pp 332–347

Rao V, Gupta I, Lokshin M, Jana S (2003) Sex workers and the cost of safe sex: the compensating differential for condom use among Calcutta prostitutes. J Dev Econ 71:585–603

Common Law System

Cristina Costantini
Department of Law, University of Perugia, Perugia, Italy

Abstract

Moving from a clear definition of the diverse and concurring meanings attributed to the syntagma "Common Law," the essay rediscovers the genealogical construction of the English legal tradition. A particular emphasis is given to the relationship between auxiliary or competing jurisdictions and, with the proper methodological approach of comparative law, to the main traits, which shape legal mentality and legal style in a different way from the canonical morphology of the Civil Law tradition.

Defining the Expression "Common Law"

The expression "Common Law" has been used with various extensions and with different, even if interrelated, meanings.

First of all, it denotes the constructive result of the act of mapping or splitting the worldwide *nomos* (the conceptual space of the normative globe) into definite families and systems, each of them identified by its proper genealogy and its governing principles. In this perspective, "Common Law" is counterposed to "Civil Law," the first denoting the legal model which flourished in the United Kingdom (except Scotland), later transplanted into the United States of America and into the Commonwealth countries with diverse extent of local adaptation; the second naming the legal order which embraced Continental Europe. The spatial dichotomy traces its origins back to the dissimilar relation entertained by the aforesaid legal experiences with a common denominator, the Roman Law. While the Civil Law systems developed from the conscious recovery and the strategic interpretation of the stratified precepts condensed in Justinian's *Corpus Juris Civilis*, the Common Law sapiently exploited the inner potentialities offered by feudal law to resist against the hegemonic expansion of Roman texts and culture. Therefore, the Civil Law family founded its prestige on the venerable authority of Roman maxims; on the contrary, the Common Law family exalted its own uniqueness and insular authenticity by the means of a fierce immunity from Roman fascination. The relics of the ancient bipartition are kept by the other nomenclature conventionally employed to juxtapose these legal entities: on the one side, it evoked the Anglo-American legal family, which retains in the name the pride of its autochthon sources; on the other side, it made manifest the Romano-Germanic legal family, which exhibits in the epithet the satisfaction for such an illustrious descent.

Historically, the opposite recall to Roman Law molded the conceptual structure of the systems in dissimilar ways; shaped otherwise their methodological outlooks, respectively based on the assumption (in the Civil Law family) and on the rejection (in the Common law family) of the codification of laws; and justified heterogeneous techniques of legal education and legal reasoning. On this ground Common Law and Civil Law came to different choices both at an ontological and epistemological level.

In a second and more specific meaning, the expression "Common Law" is closely linked to the notion of *Equity*, in order to denote the complex structure of the English Legal system, the double soul, which gave substance and form to English Law wholly considered. In particular, "Common Law" and "Equity" designate two different bodies of rules, principles, and remedies, originally settled and administered by two separate jurisdictions, respectively by the *Curia Regis*, with its internal partitions, and by the Chancellor and the Court of Chancery.

In a third sense, "Common Law" is contrasted or opposed to "statutory law," with the aim of stressing the diverse sources of legal rules. While the Common Law is the case law, that is, the law declared or created by the Courts, statutory law is the ensemble of written laws passed by the legislature or other government agency and expressed with the requisite formalities.

The Genealogy of the English Common Law: The Jurisdictional Project

Two main visions compete for the genealogical reconstruction of the Common Law identity: one corresponds to a mythological thought supporting a legitimating project; the other represents the constitutive framework conventionally used to emplot the narratological history of English system, where a proper emphasis is given to the elements which built the specificity of Common Law and measured the distance from Continental Law.

In the first perspective, the Common Law presented itself as an organic bulk of memories and customs inherited from a "time out of mind" and handed down without solution of continuity. The immemorial law has the nature of an unwritten tradition, nourished by a tacit and original knowledge, that is close to nature and to divine law (Fortescue and Chrimes 1942; Goodrich 1990).

Therefore, the superiority of the English laws was validated even by theological and philosophical arguments.

In a different manner, the orthodox account depicts the Norman Conquest of 1066 as a catastrophe irrupted into the history of England, a revolutionary event which impressed the proper direction to the future course of English Law. The pristine foundations of the Common Law are underestimated, if not concealed, and William the Conqueror, Duke of Normandy, is considered as the illuminate king responsible for the creation of the political and social context in which the principal institutions of a distinguished system could be developed (Baker 2000, 2002; Milsom 1969; Plucknett 1956).

First of all, William I imposed a particular form of feudal structure, based on a sapiently articulated association of lordship and tenures and influenced by the very conception of liegeancy. The King was proclaimed as the supreme Lord of the landed property, and as a consequence, this assertion prevented the introduction of allodial estates, over which one could exercise full and unrestricted ownership. At that time the European *nomos* was split into two opposite models, into two ideological visions of government and society competing with each other: on the one side, there was the Roman model with its clear demarcation between the concepts of *imperium* (public power) and *dominium* (private ownership); on the other, there was the feudal model with its confusing mixture of the public and the private (Samuel 2013).

Another relevant aspect of William's policy was administrative centralization, also extended to law. Enhancing the medieval conception of sovereignty, the King was the pinnacle atop the feudal hierarchy and retained both legislative power and jurisdictional capacity: he was Lord Tenant in Chief and, metaphorically, the fountain of justice, supervising the declaration of new rules and the dispensation of new remedies. The English Common Law could not originate and exist, as we actually know it, if a corps of advisers and courtiers, named *Curia Regis*, wasn't established in all its functions. This council embodied the center of royal administration: initially it was undivided and peripatetic, following the King during his itinerant circuits across the realm; later it was transformed into a fragmented and mainly stable organ, which began to sit regularly at Westminster. The internal evolution of the Curia Regis led to the formation of the three royal courts which molded the forms and the contents of a law common throughout the kingdom.

The Exchequer was the first department to be deposited, dealing with finance and taxation. Another division was settled to examine the petitions that affected the King's interests, because of their nature heard and discussed *Coram Rege*, at the presence of the Monarch. It was the King's Bench with jurisdiction over issues recognized as "public" in their orientation, such as those nowadays included within criminal and administrative law. In order to grant a secure and constant justice to all the litigants, a historical compromise was reached in 1215, inserting in the text of Magna Carta a specific clause, according to which the "common pleas" – that is to say all the suits in which the King had no interest – would be heard by a permanent body of judges set in Westminster. As a corollary, a new court originated, which decided not *coram rege* but *in banco*, known as the *Common Bench* or as the *Court of Common Pleas* (Brand 1992).

By the fourteenth up to the seventeenth century, the Common Law was a matter of three royal courts competing for litigation (Samuel 2013). The English Common Law developed and flourished as a jurisdictional project.

The consciousness of unicity and originality that marked the Common Law Tradition was supported even by the specificity of procedural technicalities required for the effective functioning of the Royal Courts.

The first characteristic was the systems of writs (Maitland et al. 1936). The plaintiff who wished to start a lawsuit before one of the Royal Courts had to purchase a writ from the Chancery section of Curia Regis, a sort of granted permission authorizing the commencement of the proceedings. This was due to the fact that the Courts in Westminster, before becoming the ordinary and regular courts of law, had an exceptional jurisdiction, allowed by royal favor; as a consequence, litigants

formally had not the right to go to the royal courts, but they needed a sort of pass to benefit from that kind of justice for which they had paid. In its materiality the writ was a strip of parchment usually written in Latin and sealed with the Great Seal; in its juridical consistency it was an order to do justice based on a definite and compelling formula, composed by selected words, apt both to introduce a particular type of action and to settle a likewise specific procedure. From the beginning of the thirteenth century they were collected and reported in a proper book entitled the *Register Brevium*. By the middle of the same century, as a result of their exponential growth, a political and juridical contention arose among the King, the Feudal Lords, and the common law judges. The King claimed to preserve the reserve of justice he bare by virtue of his paramount sovereignty; the Feudal Lords craved to avoid the infringement of their signorial jurisdiction; the Common Law judges aspired to strengthen their authority even though they were overwhelmed by litigation. A temporary composition of these tensions was reached in 1258 with the Provisions of Oxford, which fixed the orthodoxy of the common law system, insofar as no unprecedented writs could be issued by Chancery without the consent of the King's Council. This form of institutional conciliation generated order and certainty but imposed an asphyxial fixity, a stagnant and bogged immobility. The subsequent Statute of Westminster II allowed a certain, albeit limited, openness through the recognition of "*writs in consimili casu,*" so to make justiciable all those instances presenting a great similarity with others for which the writ was already dispensed. It should be emphasized that the choice of the correct writ was not a mere formality. First of all, an improper selection among the suitable writs undermined the whole procedure, while the absence of an appropriate writ was equivalent to the absence of a legal remedy; secondly, the internal classification of writs was at the basis of the legal taxonomy of actions and, in course of time, represented the structural framework of the substantive outlines of the common law.

Moreover, from the origins onward, the system of writs fashioned the common law mentality in a very different manner from the civil law way of reasoning and arguing. The common lawyers privileged analogy and factual appreciation, since the concrete facts of a dispute had to be compared with the models of factual situations already sanctioned in the *Register Brevium*. The civil lawyers privileged logic deduction, since the proper solution of a juridical issue had to be derived from a coherent complex of superior and outstanding principles.

The second and most important procedural feature was the presence of the jury in the course of the ordinary trial. Firstly used in criminal procedure as a means of evidence, the jury was subsequently imported into private lawsuits as a sworn body of lay people convened to render an impartial verdict and to judge on the facts of the controversy.

Common Law and Equity: A Mutable Relationship

The growth of the Common Law marked the victory of a centralized authority over a cluster of diffuse and competing powers. Nevertheless the balance achieved among the King, the Feudal Lords, and the Common Lawyers ultimately sacrificed the creative progress of the system of law. The English Common law froze and arrested; historically it was not too far from a shocking paralysis. The apparent involution was also made worse by the emergence of grave defects which undermined the efficiency of the Royal Courts and seriously compromised their popularity. Moreover, the set of remedies created and administered by the central Courts became progressively inadequate and irrespective of the new or increased exigencies. In particular, the common law courts were not technically equipped to grant personal remedies, that is, to order a party to do or not to do something, the only remedies dispensed being monetary remedies (debt or damages). On the other hand, if the facts of the claim could not be adapted to the formula of an established writ, the search for justice was utterly denied and the trial could not commence. The lack of a proper system of appeal courts and a mounting set of

judicial vices (among which corruption and delay) completed a not comforting framework.

However, in spite of everything, the English law was able to replenish itself from the inside. The most urgent need was to overcome the strictness of the formulary system, which could preclude the effective access to the common law courts. The historical escape resided in the overriding and residuary power of the king to dispense justice outside the regular system. By the end of the thirteenth century and during the first half of the fourteenth, the unsatisfied claimant came to address a specific petition (bill) to the King in piteous terms, asking for his grace and mercy to make manifest in respect of some complaint. Initially the King examined these formal requests directly, by himself, but due to the considerable increase of their number in times, he began to pass the petitions to the Chancellor. As it has brilliantly pointed out, the Chancellor was in direct connection with all the parts of the constitution (Holdsworth 1966). He was the secretary of state of all departments; to him was entrusted the Great Seal by which all the acts of state and royal commands are authenticated; as the head of the Chancery, formerly conceived as an administrative department, he had also the power to draw and seal the same royal writs necessary to start legal proceedings before the Courts of Common Law. Moreover, most early Chancellors were ecclesiastics, keepers of the King's conscience. When the petitions were passed from the King to the Chancellor, he formerly decided on behalf of the monarch, admitting "merciful exceptions" to general law with the aim to ensure that the King's conscience was right before God (Watt 2008). By the end of the fourteenth century, the Chancellor decided in his own name and on his own authority. Moving from his ecclesiastic affiliation, the Chancellor exercised a transcendent form of justice, beyond the common law jurisdiction and based on an innovative mixture of canon precepts, Christian discretion, and Roman Law. The procedure was different if compared with that enacted by the Royal Courts of Westminster and integrally devoted to amend the presumably guilty conduct of the defendant. In particular, all the actions were commenced by an informal complaint, in order to make unnecessary the selection of a correct writ; the pleading was conducted in English (not in Latin, nor in French); there was no jury; the final judgment was expressed in the form of a decree, more precisely in a decree of injunction or in a decree of specific performance; the Chancellor decided questions of fact as well as issues of law. One of the most important features of the new procedure was the proper form of the act whereby it starts, called *writ of subpoena*, which, in spite of its name, was something other than the old writs enacted in Common Law Courts. While the latter explicitly mentioned the cause of action against the defendant, the writ of subpoena was limited to command the physical presence of the litigant before the Chancellor – upon pain of forfeiting a sum of money – in order to answer to the complaints made against him by the plaintiff, without revealing the specific reasons led at the basis of the claim.

This was the process which led to the formation of a separate corpus of principles, remedies, and rules, autonomous from the common law, strictly considered. A line of demarcation crossed the former and undivided space of jurisdiction: the proper domain of the Courts of Westminster was disjointed from the sphere of the Chancellor, newly rediscovered (Maitland et al. 2011).

From a constitutional point of view, the same process made complex the originally unitary vision both of the Chancellor and of the so-called Curia Cancellariae, the staff of clerks over which he presided. In course of time, it was possible to recognize two sides of each of the aforesaid bodies. Looking at the figure of the Chancellor, on the one hand he had the power to seal the writs needed to bring in court a common law action, but in this guise he didn't act as a judge, he didn't hear the polemical arguments of the parties, insofar as he simply granted a writ on the basis of the plaintiff's claim, leaving to the three Royal Courts the decision on the conformity of the writ to the law of the land. On the other hand, the Chancellor gradually exercised an extraordinary form of justice, not to supersede or to contradict the principles expressed within the boundaries of the common law jurisdiction, but to mitigate the excessive rigor of the Courts and to adequate remedies to the new substantial needs.

Conversely, looking at the "Curia Cancellariae," it was possible to detach the growing of a Common Law side and an Equity side (as it was called at the end of the formative period), or a Latin side and an English side, giving relevance to the official language used in the course of the different procedures. The Latin side was requested to examine that petition which concerned the person of the King, whenever justice was demanded against the sovereign. The proceedings were enrolled in Latin and developed in a very similar manner to that followed in the three courts of law. The English side gave rise to the new form of justice, we have discussed above, originally perceived as ancillary to the common law system.

The history of this other source of English legal system was marked by progressive metamorphoses, which affected both its nature, function, and appraisal and the subtle relationship with the common law, properly considered. As a result, the whole morphology of English law took different appearances across epochs and times.

In the first direction, it's possible to detach a transformation with respect to the same structure of the organ that administered the new jurisdiction: gradually the Lord Chancellor ceased to be an individual taking decisions and became the Chief of a Court rendering justice in the name of the King. At first this Court was known as the "Court of conscience" and lately was designed as the main Court of Equity jurisdiction, using a concept – that of equity – with a polymorphous cultural heritage, as it was located at the intersection of Greek, Roman, and Christian traditions. In this perspective the English legal system institutionalized, in the proper form of a permanent jurisdiction, what in other legal experiences remained a theoretical or philosophical concept with particular and limited transpositions in the juridical domain.

In the second direction, Common Law and Equity grew up in a changeable relationship. As it has been noted, in the first genealogical period Equity was structured and recognized as a kind of supplementary justice and law: it was not a self-sufficient system, but at every point it presupposed the existence of common law; it was auxiliary, accessory, supplemental; it came to rectify the severe asperity of common law, with a discretional appreciation of the particularities embedded in single cases. Profound evolutions occurred in the course of the sixteenth century, when Equity jurisdiction appeared to be settled and consolidated. In this period, in fact, three main factors concurred to transmute the primordial aspects of Equity. First of all, Cardinal Wolsey was the last ecclesiastical Chancellor; from this time onward laymen, and eventually even great lawyers, were designated as Lord Chancellors of England. Consequently, Equity had to manage the secularization of its proper sources, the collapse of its theological foundation, and the possible collision with common law rules, which could be indirectly transposed by the new figure of Chancellor. The device used for the preservation of its autonomy marked the second renewed feature of the jurisdiction: to outlast, Equity came to adopt the same rhetoric of Common Law, putting the observance of precedents and codified rules before the consideration of the specificities of individual cases. Finally, both Tudors' and Stuarts' dynasties manipulated and adapted the same vision of Equity in order to achieve political goals (Costantini 2008). In this perspective the original nature of Equity as an extraordinary form of justice was converted into an *instrumentum regni*, into a means used to justify and support royal prerogatives against the restrictions imposed by the Law administered by the central Courts. To that end, new Courts – other than the historical Chancery Court – apt to dispense equitable remedies, were instituted: the Court of Requests, by late 1530 competent in civil matters for poor petitioners seeking relief for minor legal issues, and the Star Chamber, created by Henry VII in 1487 out of one of the traditional functions of King's Council and composed by the same members of the Privy Council when they were dealing with criminal prosecutions. The sphere of competence of the Star Chamber was progressively and strategically extended: from an agency of control and social discipline, it was transformed into an instrument of absolutism, as well as into a dreadful thread to those liberties guaranteed by common law. The Court inflicted sanctions of various nature and

intensity, from the monarch's displeasure, passing through fines and imprisonment, to conclude with corporal punishments. The procedure was a further adaptation of canon law and differed in many regards from the common law one. The effectiveness of these newborn Courts was evidence and measure of the respect with which the authority of the Sovereign was regarded.

The institutional framework, refashioned according to royal will and aspirations, laid the foundations for a parallel alteration of the peaceful relationship between Common Law and Equity jurisdictions. The harmony gave way to discord, tension, and frictions. The conflict patently arose in 1616, in James I's days, as a fierce battle among strong personalities (Simpson 1984). The leading role in the dispute was acted, on the one side, by Lord Ellesmere, in his quality of Chancellor and able common lawyer, and on the other, by the Lord Chief Justice, sir Edward Coke, the best active proponent of the reasons of the common law. The main object of contention was a recurrent practice of the Chancery Court, namely, hearing a case in Chancery after judgment had been given at common law. Obviously, the Common Law Courts were jealous of their powers and believed that this separate judgment – within the boundaries of a separate jurisdiction – was a kind of irregular appeal, contrary to statute. Therefore they made use of legal devices, and especially of writs of habeas corpus and prohibition, to fight against the illegal invasions and violations of the Prerogative Courts. Finally, given the institutional relevance, the quarrel was referred to the King, James I Stuart, in 1616. At this moment Lord Ellesmere, Francis Bacon, and the Duke of Buckingham worked together as the historical engineers of Coke's downfall (Baker 2002). Coke was dismissed from office and James I stated by decree that in case of conflict between Common Law and Equity, the rules of Equity would prevail. After Ellesmere's death, the relation between the two jurisdictions was reassessed in a cordial mood.

Another form of reversal should impress the course of English legal history. It was no longer related to the correlative dynamics of rival jurisdictions, but it concerned the proper nature and structure of Equity. What, at first, was a device apt to soften the asperities of Common Law now hardened into law. It was an historical irony: the system of remedies and principles, which, in the past, concretely corrected the deficiency and the inadequacy of Common Law, finally was pervaded by even worse defects. During the seventeenth and eighteenth century the Court of Chancery was transformed from the elected place of relief and comfort to the proper locus of anguish and despair. All the possible aberrations were collected in Chancery: the procedure had become complex and cumbersome; the mass of documents and procedural acts were elephantine; the length of judicial processes had definitely sacrificed the instance of justice.

This phase of stagnation and involution continued up to the nineteenth century, the age of the great reforms, which led the foundation of the renewed system of jurisdiction introduced by the Judicature Acts in 1873–1875. This statute abolished the old Courts and established three levels of central justice: the first was the High Court, the second consisted of the Court of Appeal, and at the third was posed the Judicial Section of the House of Lords. The High Court historically derived from the final embodiment of the old courts of law and is internally structured into three divisions, the Queen's Bench Division (QBD), the Chancery Division (ChD), and the Family Division. It's important to underline that the Victorian legislation caused the procedural fusion of the old rules (originally divided into the two bodies of Common Law and Equity); the specialization of the three divisions is only a matter of convenience, insofar as all the judges are empowered to administer both Law and Equity. According to the latest reform, the appeal body of the House of Lords has been transformed into a new Supreme Court, with an independent seat from that of the House of Lords.

Mentality and Style Within the Common Law Tradition

From a comparatistic point of view, legal traditions can be defined, distinguished, and identified

assuming legal mentality and legal style as proper markers, or demarcation devices. The concept of mentality relates to the complex of elements pertaining thinking, discourse, narrative, symbolic, and social practices shared or recognized by a given community and characterized by persistence in time. The idea of style describes the forms and the perceptive qualities of a legal system; it is a replication of patterning that derives from a series of choices made within some set of constraints (Lang 1987).

The Common Law tradition is structured on the basis of a specific legal mentality and had elaborated an original legal style, which concurs to justify the differences between Common Law and Civil Law with regard to four main aspects: legal education, form of judgment, legal taxonomy, and epistemological attitude to law.

First of all, as a matter of fact, there is a close relationship between any system of law and the professionals, the experts who operate it (Dawson 1968). Historically, the peculiarity of English Common Law Tradition and its proper resistance to continental influences rest on the spaces and methods of legal education, on the subtle link between Bench and Bar, on the internal organization of legal profession. Traditionally, common lawyers were formed by practitioners, not by law professors, into the elitist space of the Inns of Court, not into the broader dimension of the Universities, where only Roman Law and Canon Law could be taught and transmitted. Change came only in the second half of the nineteenth century, when the academic study of law was established.

Moreover, by the end of the thirteenth century it had become a general custom that judges of the central courts could only be appointed from the professional bar. This trait came to distinguish the English legal tradition, assuring the peculiar strength of its professional elite.

Nowadays the legal profession in England and Wales is divided into two branches: those of barristers and solicitors. At the beginnings of their institutions each of them retains a proper monopoly (court representation for the barristers and conveyancing for the solicitors), but the reforms passed during the 1990s abraded this pristine privileges.

Binding force of judicial decisions represents the second marker of Common Law tradition (Lasser 2004). While on the Continent the internal coherence of the legal systems was granted by the act of normative codification, in England the notorious spirit of anticodification urged to find out another juridical device apt to achieve the same goal. The doctrine of precedent (or the doctrine of *stare decisis*), by which the Courts are obliged to respect their prior decisions, was affirmed during the nineteenth century, after the reorganization of the hierarchical structure of central jurisdiction and the introduction of an official system of law reporting. The effects of these principles occur both vertically (the decisions of a superior court bind all the inferior courts) and horizontally (a court is bound to follow its own precedents unless there is a strong reason to not do so). Arguing that a too rigid adherence to precedent may lead to injustice in a particular case and also unduly restrict the proper development of the law, with the Practice Direction of 1966, the Law Lords of the English House of Lords decided to depart from a previous decision when it appears right to do so. It is important to stress that what is binding is the only *ratio decidendi*, the rule of law heated by the judge as a necessary step in reaching the conclusions (Cross and Harris 1991), to be ascertained by an analysis of the material facts of the case, and to be distinguished from the *obiter dicta*, that is, things said by the way. The doctrine of precedent comes to shape legal reasoning in a proper manner, giving a specific emphasis to legal argumentation and to the precise appreciation of facts.

The third main difference between Common Law and Civil Law is the internal taxonomy. While in the Civil Law systems the principal dichotomy is posed between substantive and procedural law, and then, within substantive law, between private and public law, in the Common Law tradition the formative history of jurisdictions has prevented from the creation of rigid distinctions. The rules are agglutinated around three blocks – persons, things, and remedies – and, even after the procedural fusion caused by the Judicature Acts, maintain the mark impressed at their origin, so to be recognized as the

expressions respectively of the Common Law or of the Equity system of justice.

Cross-References

► Civil Law System

References

Baker JH (2000) The common law tradition: lawyers, books, and the law. Hambledon Press, London
Baker JH (2002) An introduction to English legal history, 4th edn. Butterworths, London
Brand P (1992) The making of the common law. Hambledon Press, London
Costantini C (2008) Equity breaking out: politics as justice. Pólemos 1:9–20
Cross R, Harris J (1991) Precedents in English law. Oxford University Press, Oxford
Dawson JP (1968) The oracles of the law. University of Michigan Law School, Ann Arbor
Fortescue J, Chrimes SB (1942) De laudibus legum anglie. University Press, Cambridge
Goodrich P (1990) Languages of law: from logics of memory to nomadic masks. Weidenfeld and Nicolson, London
Holdsworth WS (1966) A history of English law. Methuen, London
Lasser M (2004) Judicial deliberations: a comparative analysis of judicial transparency and legitimacy. Oxford University Press, Oxford
Maitland FW, Chaytor AH, Whittaker WJ (1936) The forms of action at common law: a course of lectures. Cambridge University Press, Cambridge
Maitland FW, Chaytor AH, Whittaker WJ (2011) Equity: a course of lectures. Cambridge University Press, Cambridge
Milsom SFC (1969) Historical foundations of the common law. Butterworths, London
Lang B (Ed) (1987) The concept of style. Coneee University Press, Ithaca
Plucknett TFT (1956) A concise history of the common law. Little Brown, Boston
Samuel G (2013) A short introduction to the common law. Edward Elgar, Cheltenham UK
Simpson AWB (1984) Biographical dictionary of the common law. Butterworths, London
Watt G (2008) Equity & trusts law. Oxford University Press, Oxford

Common Sense

► Rationality

Commons, Anticommons, and Semicommons

Enrico Bertacchini
Department of Economics and Statistics "Cognetti de Martiis", University of Torino, Torino, Italy

Abstract

The notions of commons, anticommons, and semicommons are presented here to highlight their connections concerning how forms of ownership beyond the classical boundaries of private property affect the management of resources. While the three concepts have been presented as expressing specific dilemmas for the management of the resources or distinct property regimes, they may be seen as components of a unified interpretative framework which recognizes resources as collection of multiple attributes and addresses the complexity of mixed property regimes by studying the interaction of common and private uses.

Definition

A semicommons exists when the management of a resource is characterized by the coexistence of both common and private uses. Because a resource may be comprised of a bundle of attributes, which can be put to various productive uses, the introduction of semicommons highlights how different property regimes may coexist to govern the simultaneous uses of the resource. At the same time, a semicommons expresses a tragedy, because it induces problems of strategic behavior by agents in governing the externalities which emerge in the interaction between common and private uses. The semicommons therefore relates to and extends the notions of commons and anticommons in understanding how the allocation of property rights affects the management of resources.

Introduction

Commons, anticommons, and semicommons refer to three notions property law scholars have introduced to explain and analyze how forms of ownership beyond the classical boundaries of private property affect the management of resources and cope with creating incentives and dilemmas for multiple owners.

As the three concepts have been elaborated in different and subsequent periods, they reflect an evolution in the appreciation by property theory scholars of the complexity of governing resource systems through property regimes. On one hand, commons and anticommons pose two symmetric dilemmas and reflect the tension between too many use privileges and overfragmentation of private interests. On the other hand, the semicommons, with its dynamic interaction between private and common property over the same resource, offers new insights as to conditions in which mixed property regimes emerge and fragmentation solutions are favored to avoid strategic behavior in multiple uses.

This entry provides an overview of the main connections between the three concepts with a particular focus on explaining the most recent notion of the semicommons. The insights drawn from the scholarly contributions show that commons, anticommons, and semicommons may be seen as components of a comprehensive interpretative framework toward a better understanding of how resource systems are regarded as collections of multiple attributes and accommodate multiple uses that are most efficiently pursued at different scales, whether simultaneously or over time.

Commons and Anticommons Revisited

In studying common property regimes, scholars have long identified the so-called tragedy of the commons, which occurs when open-access or common ownership resources become overexploited due to the uncoordinated actions of the common right holders (Gordon 1954; Hardin 1968; Cheung 1970). The debate surrounding the commons centered on the institutional mechanisms to avoid such tragedy (Ostrom 1990) and on the optimality of common versus private property regimes in managing physical nonexclusive resources.

More recently, scholars have turned their attention to anticommons phenomena. In this context, concurrent controls on entry over a common resource exercised by individual co-owners acting under conditions of individualistic competition and exclusion rights will be exercised even when the use of the common resource by one party could yield net social benefits.

While law and economics scholars have long recognized how multiple vetoes in contractual relations generate holdup problems which could produce inefficiency and underinvestment (Klein et al. 1978; Williamson 1979; Hart and Moore 1988), Heller (1998) has been the one who has initially popularized the definition of the anticommons and made more explicit this dilemma in property theory by drawing observation from the development in post-Communist Europe, where many buildings remain empty while numerous kiosks occupy the streets.

Introducing the anticommons has allowed to unveil similarities and peculiar symmetric relationships with the commons, which in turn helps analyzing the optimality of different property regimes.

According to formalized models presented by some authors (Buchanan and Yoon 2000; Schulz et al. 2002), commons and anticommons lead to symmetric tragic outcomes. Both the commons and the anticommons tragedies feature self-interested choices that are collectively suboptimal. However, in common property regimes, too many use privileges lead to overexploitation of the common resource, while in the latter fragmentation in property and too many exclusion rights lead to underuse of the asset. In this perspective, Heller (1998, 1999) points out how commons and anticommons may be intended as property regimes which provide a useful framework to define the economic implications of property law for the management of resources. The choice of property regimes may in fact be seen as a continuum along commons, private property, and anticommons regimes. In this model the main

challenge is to set the boundaries of private property by taking into account two diverging forces. On the one hand, it is necessary to consider the trade-off that occurs in the so-called tragedy of the commons between the benefits derived from large-scale operations and the costs of overuse. On the other hand, property fragmentation may generate the familiar anticommons result of underexploitation.

While acknowledging a symmetric relationship between commons and anticommons, some scholars have nonetheless warned from regarding the anticommons as a new distinct type of property regime. For example, Lueck and Miceli (2007) argue that anticommons should more properly be seen as an "open-access" investment problem: a resource, such as land or apartments, can be overused by multiple owners, but the investment on the asset can be simultaneously suboptimal because of the lack of unanimous agreement by the multiple owners. In a similar vein and focusing on a game theoretical approach, Fennell (2011) suggests that the most promising and general approach to assess commons and anticommons is to consider the distinct strategic behavior underlying the two situations. Whereas the commons tragedy follows the strategic pattern of the prisoner's dilemma, the anticommons often resembles the strategic game of chicken.

Commons and anticommons may be therefore seen as useful metaphors for understanding the misalignment of incentives of multiple owners who wish to use a common resource. Much of the confusion concerning the definition of commons and anticommons either as dilemmas in the management of common resources or as distinct property regimes stems from the fact that most of the studies on property regimes assume that each underlying resource possesses one unique attribute that allows for a single productive activity to take place. To the contrary, however, a resource system may be comprised of a bundle of useful attributes, which can be put to various productive uses and at different scales of operation (Barzel 1982; Lancaster 1966; Smith 2002). Considering the familiar example of the grazing field used to explain the tragedy of the commons, Fennell (2011) points out how the problem is not only related to the fact that grazing is pursued at a large scale and on commonly owned ground, but also to the fact that the commonly owned attributes interact with other privately owned inputs, such as cattle. The overuse of the commons emerges from the appropriation of benefit to the privately owned attribute of the resource system. If only land were privatized, overgrazing would be avoided, but the misalignment of incentives and the overuse would affect other attributes or inputs still experienced in common, such as water. As a result, the intrinsic dilemma of the commons is more properly due to a problem of efficient specification of what attributes of a resource system are held in private or common ownership and how those property regimes interact in the productive use. In a similar vein, in the anticommons case, fragmentation of property over a resource may be seen as a private individual property arrangement held by distinct owners as long as some attributes of the resource fail to be available for potential productive use at a greater scale of operation due to assembly problem of property rights.

Introducing the Semicommons

The recognition that resources hold bundle of useful attributes implies that different property regimes may coexist to govern the simultaneous uses of the resource and add new insights in the analysis of the complexity of mixed ownership regimes.

In this context, Smith (2000) has developed the notion of the semicommons, a property regime in which one attribute of a resource is privately owned, while another is in common ownership, and the two potentially interact. The now classic example of the semicommons is taken from medieval open fields. In such fields, peasants had exclusive property rights in strips of land for purposes of growing grain but shared these land strips with other farmers in one large grazing commons during fallow seasons. A semicommons property regime enables the various right holders to benefit from the multiple uses of the resource. First, semicommons owners obtain the advantage of scale

economies. This is demonstrated in the medieval field example, as the peasants hold the resource in common for grazing. Secondly, semicommons harness private incentives by allocating privately owned rights on the resource. Interestingly, a semicommons property regime may induce problems of strategic behavior that go beyond the familiar incentives of overuse that exist in a common property regime.

Similar to the tragedy of the commons, semicommons members may overuse the common resource to the extent that they internalize merely a fraction of the cost of their actions. Additionally, however, because of the interaction between common and private uses, owners will attempt to distribute benefits to their own part of the commons and steer bads to the other parts of the assets. In the open field example, peasants have incentives to allocate the benefits of manure onto the part of the commons that they privately own during fallow season and to further steer the damage of trampling onto the land of others (Smith 2000). Institutional solutions are needed to moderate such strategic behavior in a semicommons setting. In the example of open fields, the scattering of privately owned strips reduces strategic behavior among farmers. By randomly dispersing the privately held parcels, the altered borders of the land make the costs of engaging in strategic behavior prohibitive.

Semicommons Between Commons and Anticommons

Because of its peculiar characteristics, the semicommons holds fundamental relationships with both commons and anticommons arrangements.

From an analytical viewpoint, semicommons has been related to the commons using the latter as a reference benchmark to analyze the economic benefits and cost of the two ownership structures. On one hand, semicommon property allows operation on a larger scale than common property by enabling both common and private uses. If adding one productive activity leads to a more efficient use of the resource (i.e., by increasing total output), the semicommons would be a superior solution instead of a pure common or private property regime where only one productive use is performed. On the other hand, it poses strategic problems that may well go beyond the familiar incentive of overuse in a commons (Smith 2000). In particular the uncoordinated behavior by agents in a pure commons is likely to generate "equally shared" externalities among agents. By contrast, in a semicommons, actions taken while the resource is in common use generate positive or negative external effects on the privately owned parcels. For this reason, agents in a semicommons will try to act strategically by trying to distribute in unequal share to their privately owned attributes the negative or positive externalities arising from their common use activities. As a result, the interaction between private and common uses induces strategic behavior that may undermine the gains provided by multiple uses of the resource and potentially leads to a lower value of using the resource in a mixed property regime. Considering negative externalities generated by common use to the privately owned attributes of the resource, Bertacchini et al. (2009) have confirmed through a formalized model the more tragic outcome of semicommons property arrangements as compared to the stylized commons, with an increased overexploitation of the resource by the common use activity. However, it is not still clear whether the same conclusion applies in the presence of positive spillovers from common use to privately owned attributes use or vice versa.

As for the relationship between semicommons and anticommons, this may be found in the role of scattering strategy in the privately owned attribute of the resource.

According to Smith (2000), scattering reduces the opportunities for strategic behavior. However, it may also reduce the efficiency of crop production, since the privately owned attribute is fragmented to a scale of operation that might be below the efficiency standard. This highlights the trade-off that agents face when they choose to implement a scattering strategy. On the one hand, agents need to fragment the privately owned attribute enough so as to increase the costs of strategic behavior and maintain the

multiple use of the resource. On the other hand, agents seek to minimize the losses that derive from the efficiency reduction in the production activity pursued with private use.

As noted by Bertacchini et al. (2009), a rule of scattering allows agents to contract into an anticommons regime in order to create a mix of private and common property. The use of scattering illustrates in fact a potential virtuous linkage between commons and anticommons property regimes. A number of scholars have suggested that an anticommons regime is a desirable allocation of property rights when nonuse of the resource is the preferred equilibrium, such as in the context of conservation management or environmental preservation of resources (Mahoney 2002; Parisi et al. 2005). Analysis of semicommons ownership reveals an additional benefit of anticommons property arrangements. When the mix of private and common uses generates strategic behavior, fragmenting a resource for one activity can thus be useful to achieve efficiency with regard to other activities. For instance, a semicommons can introduce a benevolent "comedy" of the anticommons by fragmenting property rights in an attribute of the resource in order to realize scale benefits of the resource's other attributes.

More generally, the introduction of the semicommons opens the door to a broader analysis of property regimes and resources that involve different scales of use.

According to Fennell (2011), the semicommons is thus less a distinctive property type than a frame through which to view existing or proposed arrangements that involve activities at different scales, whether simultaneously or over time. On this account, Smith (2000, 2005) recognizes that scattering strategies of privately owned attributes are quite common forms of boundary placement, such as proportionate holding schemes or ex ante uncertainty about the appropriation of benefits deriving from common use activities to one's holdings. In this sense, scattering is a fragmentation strategy which enables alignment of incentives by providing a more cost-effective solution as compared to maintaining an attribute of a given resource in common and establishing governance norms to monitor compliance.

Other Applications of the Semicommons

Other than the semicommons in the English open field, there is a growing literature which extends the use of this concept in other domains.

The interpretative framework of the semicommons is particularly suited to analyze property rights over fugitive resources, such as water (Smith 2008), or assets contributed to a joint venture (Smith 2005). In this latter case, an asset can be used for purposes of the joint venture, but the joint ventures may retain certain private uses and engage in strategic behavior to extract benefits for assets over which they retain some private uses and dump costs to the assets over which others have retained private uses.

More importantly, the notion of semicommons has been applied to intellectual resources, such as information and knowledge.

Because information is difficult to subject to exclusive rights and because multiple uses may derive from the non-rivalrous resource, intellectual property law is argued to establish a semicommon property regime in information goods (Heverly 2003; Frischmann 2007). In this case, intellectual property law devises a dynamic interaction between the common use and the protection of private production of intellectual resources. Although it is true that intellectual property rights partly enclose the information commons (public domain), it is equally true that they feed in many ways the information commons. This may occur, for instance, when intellectual property rights expire or in cases of research exemption in patents and fair use doctrine in copyright laws. Likewise, the information commons partly enclosed by the intellectual property rights provides information inputs for the private creation of new intellectual resources. Strategic behaviors arise also in the information semicommons and can be of two types. The illegal reproduction and distribution of privately owned information may be deemed as an improper expansion of common use because pirates

strategically distribute harms to the owners of the protected information.

To the same extent, the strengthening of exclusive rights that blocks the access to the common components of information may be seen as an expansion in protection of private use that restricts the use of the information in public domain.

A slightly different application of the semicommons in intellectual resources may be found in Reichman (2011), dealing with innovation and intellectual property regimes.

According to Reichman, the realm of industrial property law may be divided into three spheres: a commons, a semicommons, and exclusive rights (in the form of patents). The commons is characterized by free flow of scientific and technical information that are extensively government generated or funded in public research programs. At the other extreme, patents confer exclusive rights to innovators as reward for their investments in innovative endeavor and historically have been granted for truly nonobvious inventions. Between these extremes lies the main area of industrial innovative activity that does not rely on path-breaking and discontinuous inventions but rather on cumulative and routine applications of know-how to industry.

The cumulative development of know-how is seen as a semicommons because it reflects a community project that benefits from the small-scale contributions.

Indeed, small-scale innovators draw from the public domain to make improvements and enrich the public domain by generating new information that others in the technical community may exploit to their own advantage (dynamic interaction between private and common use). The cumulative additions are based on the private use of information inputs and reverse engineering practices.

Nevertheless, these additions do not attract exclusive property rights because they normally do not surpass the ability of the routine engineers who comprise the relevant technical communities.

Because the overall problem of innovators is to recoup their investment in innovation through the availability of lead time, the patent system grants legal lead time to nonobvious inventions. On the contrary, in the semicommons of incremental applications, innovators have just a natural lead time. In this case, the lead time greatly depends either on the protection of trade secrecy from unlawful misappropriation of new industrial applications or on the technological conditions that allow competitors to lawfully reverse engineer the industrial application. In summary, under the Patents-Trade Secret system, the semicommons of innovative applications of know-how favors the spread of innovation in a healthy competitive environment. At the same time it impedes single small-scale innovators to strategically behave, removing their contributions from the semicommons by means of exclusive property rights.

References

Barzel Y (1982) Measurement cost and the organization of markets. J Law Econ 25:27–48

Bertacchini E, De Mot J, Depoorter B (2009) Never two without three: commons, anticommons and semicommons. Rev Law Econ 5(1):163–176

Buchanan J, Yoon Y (2000) Symmetric tragedies: commons and anticommons property. J Law Econ 43:1–13

Cheung S (1970) The structure of a contract and the theory of a non-exclusive resource. Journal of Law and Economics 13(1): 49–70.

Fennell LA (2011) Commons, anticommons, semmicommons. In: Ayotte K, Smith HE (eds) Research handbook on the economics of property law. Edward Elgar, Cheltenham

Frischmann BM (2007) Evaluating the demsetzian trend in copyright law. Rev Law Econ 3(3):649–677

Gordon S (1954) The economic theory of a common-property resource: the fishery. J Polit Econ 62(2):124–142

Hardin G (1968) The tragedy of the commons. Science 162:1243–1248

Hart O, Moore J (1988) Incomplete contracts and renegotiation. Econometrica 56(4):755–785

Heller M (1998) The tragedy of the anticommons: property in the transition from Marx to markets. Harv Law Rev 111:621–687

Heller M (1999) The boundaries of private property. Yale Law J 108:1163–1223

Heverly RA (2003) The information semicommons. Berkeley Technol Law J 18:1127–1189

Klein B, Crawford RG, Alchian AA (1978) Vertical integration, appropriable rents, and the competitive contracting process. J Law Econ 21(2):297–326

Lancaster KJ (1966) A new approach to consumer theory. J Polit Econ 74:132–156

Lueck D, Miceli T (2007) Property law. In: Shavell S, Polinsky AM (eds) Handbook of law & economics. Elsevier, Amsterdam
Mahoney JD (2002) Perpetual restrictions on land and the problem of the future. Virginia Law Rev 88:739–775
Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge, UK
Parisi F, Depoorter B, Schulz N (2005) Duality in property: commons and anticommons. Int Rev Law Econ 25:578–591
Reichman JH (2011) How trade secrecy law generates a natural semicommons of innovative know-how. In: Dreyfuss RC, Strandburg KJ (eds) From the law and theory of trade secrecy: a handbook of contemporary research. Edward Elgar, Cheltenham, pp 185–200
Schulz N, Parisi F, Depoorter B (2002) Fragmentation in property: towards a general model. J Inst Theor Econ 158(4):594–613
Smith HE (2000) Semicommon property rights and scattering in the open fields. J Leg Stud 29:131–169
Smith HE (2002) Exclusion versus governance: two strategies for delineating property rights. J Leg Stud 31:453–487
Smith HE (2005) Governing the tele-semicommons. Yale Law J Regul 22:289–314
Smith HE (2008) Governing water: the semicommons of fluid property rights. Ariz Law Rev 50(2):445–478
Williamson OE (1979) Transaction-cost economics: the governance of contractual relations. J Law Econ 22(2):233–261

Company Liability

► Organizational Liability

Competition Policy: France

Marc Deschamps
CRESE EA3190, Université Bourgogne Franche-Comté, Besançon, France

Abstract

From a law and economics perspective, competition policy has always been a natural connection between law, economics, and politics. Antitrust is at its heart, but competition policy is a much larger topic. In our entry, we will shed some light on how competition policy is conceived and structured in modern-day France.

Introduction

The expression "competition policy" is sometimes misused as synonymous with "antitrust." This is a mistake because competition policy is the articulation between politics, law, and economics, while antitrust is generally understood to be the totality of legal and economic provisions for dealing with infringements of the competition rules (i.e., essentially abuse of dominant position, cartel, concentration). In other words, antitrust, understood in its most common sense, evacuates the political dimension and concentrates on the technical aspects. This clarification is not merely a terminological one but also makes it clear that, in the field of competition policy, it is impossible not to make political choices and that, in order to be coherent, they must be in line with all the others policies. Moreover, this distinction makes it possible to understand why it is possible, for example, to observe that, apart from the case of a different technical assessment, two countries may consider similar practices of businesses differently or why they do not have the same procedural or substantive rules. Thus, far from following the same path, countries choose their policy according to their size, their degree of openness, their growth strategy, and so on.

Moreover, because France is a member of the European Union, France's competition policy is politically and legally constructed on two bases: national law and European law (i.e., the law deriving from the Council of Europe and the law of the European Union). Under the European treaties, European Union law is an own legal order with primacy over national rights and direct effect. We will here concentrate on the law and actors at the national level.

In the remainder of this article, we will present, in turn, the essential elements relating to the genesis of French competition policy (1), its main players (2), and the structure of French competition law (3).

Origins and Development of French Competition Policy

The idea that free trade is the best way to promote prosperity emerged in France at least since Montesquieu. Yet what characterized prerevolutionary France is its corporatist organization and the existence of *jurandes*, that is to say, bodies of trade whose members held a monopoly. The spirit of competition was introduced to France in 1791 with two laws: first, that of March 2–17, known as the Allarde decree, which established the principle of freedom of commerce and industry. Further, the law of June 14–17, known as the Le Chapelier law, definitively abolished guilds. With the Penal Code of 1806, there is a first provision, Article 419, which prohibits the hoarding of commodities in order to raise prices. These remain the main elements that characterize French competition policy before the second half of the twentieth century.

France began to introduce a competition law with the ordinance of 30 June 1945 on prices. After there was: the decree of 9 August 1953 (Prohibition of Unlawful Settlement Agreements and Establishment of a Technical Commission on Cartels), the Act of 2 July 1963 (Prohibition of Abuses of Dominant Positions and Establishment of a Technical Commission on Restrictive Practices and Dominant Positions), and the Act of 19 July 1977 (creation of the Competition Commission with additional advisory functions on mergers and any competition issues).

A paradigm change took place with the ordinance of 1 December 1986, since it was on this basis that France established a genuine autonomous competition law based on the general principle of freedom of prices (Article L.410-1 and L.410-2 of the French Commercial Code). This empowerment with regard to political control is materialized by the creation of the *Conseil de la Concurrence* (Competition Council), an independent body henceforth given the power of sanctioning businesses in the event of anticompetitive practices (taking it over from the Minister in charge of the economy). The Council is controlled by the judiciary (the *Cour d'appel de Paris* and the *Cour de cassation*). The same decree extends the scope for referral and offers a procedure that better guarantees the rights of the concerned persons. Subsequently, the law of 11 December 1992 empowers the Competition Council to apply European provisions, and the law of 15 May 2001 strengthens competition law by introducing leniency and settlement procedures, raising the maximum penalties, improving international cooperation, and introducing systematic monitoring of mergers. Under the European leadership, the decree of 4 November 2004 aligns the powers of the Competition Council with those of the other European competition authorities.

The last break took place in 2008. Indeed, the law of 4 August created the Competition Authority (*Autorité de la Concurrence*, *ADLC*) and affords this Competition Authority the ability to conduct investigations of its own, the ability to self-refer in advisory matters, and the power to decide in terms of control of concentrations. Until then the Competition Council was only responsible for the analysis of the merger, and the decision remained with the minister in charge of the economy. In addition, the decree of 13 November gave the Competition Authority an investigative unit under the responsibility of a general rapporteur, as well as the capacity to issue behavioral and structural injunctions.

Actors in French Competition Policy

The French competition policy is organized around three groups of actors: political actors, the Competition Authority, and the courts.

Except for sporadic laws and regulations, political actors now exercise only a residual role. Indeed, there are essentially only three options available to the government to intervene in matters of competition policy. The first is the Competition, Consumer Affairs, and Prevention of Fraud Directorate General (DGCCRF), which, with 3000 agents under the minister's authority, is responsible for the competition regulation of markets by fighting unfair commercial practices between traders and anticompetitive practices. On the understanding that the DGCCRF has only the power of injunction and transaction for local practices (less than 200 million euros for all responsible companies), the ADLC retains the

power to take preliminary action and ensures that the case is dealt with in the event of refusal or nonfulfillment of the undertakings. The second possibility for the minister responsible for the economy to intervene in competition policy concerns merger control (Article L.430-7-1 of the French Commercial Code). The minister may ask for a thorough examination, and he can also finally take a decision that is contrary to that of the ADLC on grounds of general interest other than the maintenance of competition (e.g., industrial development, creation or conservation of employment). Thirdly, the minister in charge of the economy has certain prerogatives and powers with regard to practices restricting competition.

Under Article L.461-1 of the French Commercial Code, the ADLC is an independent administrative authority responsible for ensuring the free movement of competition and for supporting the competitive functioning of markets at European and international levels. It has an annual budget of around 20 million euros and employs about 200 people. It has three main components:

1. A 17-member college with a president and four vice-presidents holding full-time positions (the other members are nonpermanent and hold positions in jurisdictions, associations, or businesses); this college publishes opinions (advisory function) and decisions (decision-making function and merger control).
2. Investigative services under the authority of the general rapporteur.
3. An auditor who acts as procedural mediator at the disposal of the involved parties.

It should therefore be noted that the ADLC clearly separates the investigative acts, which are the responsibility of the general rapporteur, and the opinions and decisions, which are the responsibility of the college. Since the law of August 6, 2015, the ADLC also offers a mapping proposal to the ministries of justice and the economy for notaries, bailiffs, and auctioneers.

There are two reasons that courts have a central role in the area of competition in France. First, it is they who are in charge of the control of the ADLC: (1) in the case of anticompetitive practices, the *Cour d'appel de Paris* is the appellate court, and the *Cour de Cassation* acts as supreme court; and (2) the *Conseil d'Etat* has a decision function: in merger control, for the advisory function of ADLC, as well as for all administrative acts it takes in relation to it. Further, both the judicial and the administrative courts have to deal with certain disputes, such as those relating to restrictive practices, acts of unfair competition, and individual or collective actions for damages.

The Structure of French Competition Law

French competition law is sometimes split into two parts for pedagogical purposes: the "small competition law" and the "large competition law." It is within the first that the essence of the national specificities lies.

"Small competition law" designates the provisions of the French Commercial Code and case law relating to tariff transparency, noncompetition clauses, restriction of competition, and unfair competition. We shall briefly mention only the latter two. As early as 1945, France chose to issue rules to protect the contractor without even having to prove a breach of the market. These rules constitute what are called restrictive practices. France does not have a law defining unfair competition, which is therefore purely based on case law. Legal doctrine classically distinguishes between acts of confusion, denigration, disorganization, and parasitism.

"Large competition law" corresponds to national provisions on anticompetitive practices (cartel and abuse of dominant position) and merger control, decisions of the ADLC, and the case law of the *Cour d'appel de Paris*, the *Cour de cassation*, and the *Conseil d'Etat*. To this the European provisions, decisions, and case law relating to these areas or concerning the control of State aid must be added.

Conclusion

French competition law has followed the European and international trends and today has

a good record since, for example, its antitrust appears to conform to the highest world standards as is proved by its five-stars rating by the Global Competition Review.

Nevertheless, according to us, there remain four main challenges for French competition policy. The first is to create a global and clear structure for the restrictive practices of competition and unfair competition. Second, it would be very useful for all shareholders to have a ranking of the ADLC's decisions and opinions as it is done by the *Cour de Cassation* or the *Conseil d'Etat*. Third, opinions and decisions would make more sense if the Competition Authority would develop economic models in line with economic academic standards. Last but not least, it would be more democratic and efficient if everybody could understand who really decides the French competition *policy* and what it will be for the next years, as it is the case, for example, at the federal level for the United States or for the European Union.

Cross-References

- Abuse of Dominance
- Act-Based Sanctions
- Cartels and Collusion
- Competition Policy: France
- Conflict of Interest
- European Law
- Leniency Programs
- Merger Control

References

Autorité de la concurrence. see http://www.autoritedelaconcurrence.fr/user/index.php?lang=en. Accessed 1 Jan 2017

Code de commerce. see https://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000005634379. Accessed 1 Jan 2017

DGCCRF. see http://www.economie.gouv.fr/dgccrf/concurrence. Accessed 1 Jan 2017

Cour d'appel de Paris. see http://www.ca-paris.justice.fr/. Accessed 1 Jan 2017

Cour de cassation. see https://www.courdecassation.fr/. Accessed 1 Jan 2017

Conseil d'Etat. see http://english.conseil-etat.fr/. Accessed 1 Jan 2017

Competitive Neutrality

Régis Lanneau
Law school, CRDP, Université de Paris Nanterre, Nanterre, France

C

Definition

The concept of competitive neutrality is not yet a legal concept in most OECD countries; nevertheless, the OECD is pushing for its wider integration. Competitive neutrality is a "regulatory framework (i) within which public and private enterprises face the same set of rules and (ii) where no contact with the state brings competitive advantage to any market participant" (OECD 2009). From an economic point of view, it is providing a narrative through which most public law could be reinterpreted.

Introduction

The concept of competitive neutrality is relatively new. Its first legal occurrence dated back in 1995 in Australia's "Competition Principles Agreement." In 1996, the Commonwealth of Australia released its own "Competitive Neutrality Policy Statement." In the beginning of the 2000s, the Netherlands, Finland, and Sweden, while not necessarily using the word, developed regulations dealing with competitive neutrality problems. The European Union, through its treaties and their interpretation by the European Court of Justice, could also be considered as having developed a competitive neutrality framework. Since 2004 and a document entitled "Regulating Market Activities by Public Sector," the OECD is producing guidelines and gathering good practices regarding competitive neutrality, no less than nine reports and organized conferences around this topic. Following OECD, the UNCTAD also introduced the concept in 2014 to inquire into the practices of some developing countries (including

China, India, or Malaysia). As the OECD recognized on its website, "While the principle of competitive neutrality is gaining wide support around the world, obtaining it in practice is a much more difficult question."

Indeed, if the idea of competitive neutrality might appear fairly simple in theory, its practical aspects are far from being easy especially since the notion does not have a stable meaning and evolved through time. At first and in its narrowest conception, "Competitive neutrality (CN) requires that government business activities do not have net competitive advantages over their private sector competitors simply as a result of their public ownership" (Commonwealth of Australia 1998). In a broader approach, it is a "regulatory framework (i) within which public and private enterprises face the same set of rules and (ii) where no contact with the state brings competitive advantage to any market participant" (OECD 2009). From these definitions, this concept is addressing the question of the limits of state intervention in the economy especially when such intervention could distort competitive positions: "Competitive neutrality occurs where no entity operating in an economic market is subject to undue competitive advantages or disadvantages" (OECD 2012).

Since this concept cannot be understood without its economic rationale, it could be considered that law and economics considerations are constitutive of its foundations. From a more legal point of view, competitive neutrality could play a key function in reorganizing and unifying (through legal interpretation) different areas of law, from competition law to state aids, procurement, and tax law.

The Economic Rationale of Competitive Neutrality

For the OECD, "the main economic rationale for pursuing competitive neutrality is that it enhances allocative" (OECD 2012). Indeed, competitive neutrality is following the same logic as competition law, even if it is also possible to consider that its function is also to promote accountability of decision makers. This economic rationale is sometimes criticized for its probable consequences, reducing the size of the public sector or the range of state interventions.

Promoting Efficiency and Accountability

Competitive neutrality appears to address the issue of competition between the public sector (and especially state-owned enterprise) and the private sector businesses. Indeed, both sectors are sometimes competing to provide the same goods or services to consumers. If public sector enterprises are providing these goods and services while benefiting from competitive advantages (resulting from different regulations – including exemption from competition law – interest rates, easier access to public data, procurement contract, or taxes to mention few sources of advantages), distortions of competition might occur leading to misallocation of resources. Indeed, according to mainstream economic theory, in such a system, there are no reasons why the goods and services will be produced and provided by those who can do it most efficiently. Competitive neutrality would then be a way to maintain the integrity of the markets and thus their allocative efficiency. From a dynamic point of view, insuring the integrity of the markets is also not without consequences regarding innovation and economic development (e.g., Graham and Richardson 1995); it should also logically push public entities to operate more efficiently through the discipline provided by the market.

The other idea behind competitive neutrality is that, in a neutral environment (or in the process of achieving such an environment), it would be possible to assess the true cost of the provision of public services when these services are provided by entity which are also competing with the private sector in some branch of their activities. Democratic debate could then be based on empirical data.

Promoting Efficiency or Reducing the Size of the Public Sector and the Range of State Intervention?

Since competitive neutrality is necessarily restricting the type of governmental intervention and forcing state-owned companies to adopt "private" behaviors, some might believe that the

concept's purpose is merely to reduce state intervention and the size of the public sector. It is difficult to address this question in this entry. However, three points should be mentioned.

First, when services of general interests are not provided adequately (or at all) by private agents in a market, public authorities are not hampered to provide these services through public entities or after call for tender. Nevertheless, in such a case, the entity providing the service, especially when it is operating also in a competitive domain, should neither be over nor under compensated to avoid competitive distortions.

Second, it should be remembered, at least in the context of the European Union, that some activities are not considered as "economic" like the army, air navigation, maritime traffic, antipollution surveillance, organization, financing, and operation of prison. Moreover, this list can include other activities when they are structured in a certain way like social security or health care based on solidarity principles or public education. Lastly, certain activities, because of their size, are often not considered as threat to competitive neutrality like local museum, local hospital, or swimming pool (European Union 2011).

Third, in situation of crisis, sticking to competitive neutrality could be unproductive. Bank bailouts are often mentioned (OECD 2009).

Competitive Neutrality in Practice

Giving substance to competitive neutrality is far from being easy (OECD 2015). Indeed, this concept is much more an aspiration than a perfectly feasible reality. Two dimensions should be considered with attention: its scope and implementation principles.

The Scope of Competitive Neutrality and State's Margin of Appreciation

When addressing competitive neutrality, the first task is to identify its scope. At this level, two dimensions should be considered: its breadth and its depth.

Regarding the former, the problem is to define what economic activities are. As it has been mentioned before, it is possible to adopt an extensive conception – providing goods and services – or a more restricted one considering that these goods and services are "economic," thus reducing the scope of the market, or through the exemption of certain activities because of their size, defined by commercial turnover (Australia), market power, or capacity to distort markets. In any case, it will be required to justify exemptions from this principle. These justifications are often economic in their nature (the good or the service cannot be provided through markets), but political justifications are quite frequent (solidarity, services of general interests, sovereignty, independence, etc.).

Regarding the later, the problem is to identify both the entities subjected to competitive neutrality and the type of governmental interventions which could have impacts on competition. Indeed, according to the narrow definition of competitive neutrality, only state-owned enterprises are considered, and the purpose is merely to insure they are not benefiting from any competitive advantage because of their "public" nature. A broader conception will consider that the ownership structure should be irrelevant to competitive positions. An even broader conception will consider all economic entities and will only focus on state direct and indirect state interventions. It is at this level that the concept of competitive neutrality is the most interesting because it is forcing us to identify the type of intervention that could influence competitive positions. Not only are subsidies and compensations for public services' obligations concerned; governance structures, regulations (and especially administrative law, contract law, and labor law), taxes, cost of capital, access to public contracts, etc. are also relevant dimensions (OECD 2012; Lanneau 2016). Virtually all interventions could have consequences on competitive positions of economic agents. Only a case-by-case approach could allow the identification of these effects.

Implementation Principles

If competitive neutrality is difficult to achieve, some procedural requirements for its implementation are certainly required.

First and foremost, transparency is a key requirement. Indeed, to assess if some entities are not benefiting from a favorable treatment by the state, it is required that its intervention regarding these entities should be made in perfect transparency. This is especially true regarding the compensation for public service obligations but also for grants, loans, or other services provided by the state for the benefit of some entities. This transparency could also force some entities to maintain two separate accountabilities, one for commercial activities and the other for non-commercial activities, to facilitate the evaluation of the adequacy of compensations for public services' obligations. These principles are, for example, enacted in the transparency directive (80/723/EEC of 25 June 1980 amended several times and codified by the directive 2006/111/EC of 16 November 2006) of the European Commission. In any cases, the value of the advantage – if an advantage is identified – will be difficult to assess.

Second, some procedures should be developed to allow competitor to challenge some state's interventions, not only direct but also indirect. Indeed, since it is difficult to identify ex ante the interventions that could lead to offer some competitive advantages, it is required to decentralize the possibility to identify "problematic" interventions. Of course, some exemptions by categories could be enacted to reduce enforcement costs, but most of these categories cannot be ascertained ex ante. For some specific acts, some ex ante procedure could also be created to avoid the consequences of competitive distortion.

Conclusion

Competitive neutrality could offer a real opportunity to address, under a common framework, different areas of law or regulations which are often considered as separated. Moreover, it would offer the opportunity to compare national practices because of its functional dimension. If the concept has not yet entered in many national legislations, its potentiality is difficult to deny, and law and economics could find, in this topic, a fantastic field of study.

Cross-References

- ▶ Political Competition
- ▶ Public Choice: The Virginia School
- ▶ Public Goods
- ▶ Public Interest
- ▶ Regulatory Impact Assessment
- ▶ State Aids and Subsidies
- ▶ State-Owned Enterprises

References

Commonwealth of Australia (1998) Commonwealth competitive neutrality guidelines for managers. Available online: http://archive.treasury.gov.au/documents/274/PDF/cnguide.pdf

European Union (2011) Communication from the Commission on the application of the European Union State aid rules to compensation granted for the provision of services of general economic interest. OJEU C 8 of 11.1.2012, http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:008:0004:0014:EN:PDF

Graham E, Richardson D (1995) Competition policies for the global economy. Columbia University Press, New York

Lanneau R (2016) La Neutralité Concurrentielle, Nouvelle Boussole du Droit (Public) Economique. Droit administratif, 2016 no. 1

OECD (2004) Regulating market activities by public sector, DAF/COMP(2004)36. Available online: http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DAF/COMP(2004)36&docLanguage=En

OECD (2009) Policy roundtables, state owned enterprises and the principle of competitive neutrality. Available online: http://www.oecd.org/daf/ca/corporategovernanceofstate-ownedenterprises/50251005.pdf

OECD (2012) Competitive neutrality: maintaining a level playing field between public and private business. Available online: http://www.oecd.org/daf/ca/corporategovernanceofstate-ownedenterprises/50302961.pdf

OECD (2015) Roundtable on competitive neutrality in competition enforcement – note by the European Union, DAF/COMP/WD(2015)31. Available online: http://ec.europa.eu/competition/international/multilateral/2015_june_competitive_neutrality_en.pdf

UNCTAD (2014) UNCTAD research partnership platform: competitive neutrality and its application in selected developing countries. Available online: http://unctad.org/en/PublicationsLibrary/ditcclpmisc2014d1_en.pdf

Concentrated Ownership

Caspar Rose
Copenhagen Business School, Copenhagen, Denmark

Abstract

This entry summarizes the main theoretical contributions and empirical findings in relation to concentrated ownership from a law and economics perspective. The various forms of concentrated ownership are described as well as analyzed from the perspective of the legal protection of investors, especially minority shareholders. Concentrated ownership is associated with benefits and costs. Concentrated ownership may reduce agency costs by increased monitoring of top management. However, concentrated ownership may also provide dominating owners with private benefits of control.

Synonyms

Blockholders; Majority control

Mitigating Agency Costs Through Concentrated Ownership

Modern listed firms are characterized by having a diffuse ownership structure with a profound separation between ownership and control. This is contrary to the heavily concentrated ownership structure that dominated business life some centuries ago where a patriarchal management structure – often relying on a family that maintained control over the firm's business and operations. However, as production became more capital intense due to the technological expansion in the eighteenth century, there was a need to expand ownership in order to attract sufficient capital. In some countries, e.g., in Northern Europe, the founder or his/her descendants often maintained control of the firm, since they held the shares with superior voting rights, whereas outside suppliers of equity held shares with inferior voting rights.

The separation between management and ownership in modern firms creates some potential conflicts of interests between management and shareholders. Agency costs are generated due to the separation of ownership as investors are not able to monitor management without incurring costs (cf. the free rider problem). This is foreseen by the external providers of capital, so even if management has identified projects with positive net present value, it may find it difficult to convince capital suppliers to invest their wealth in the firm (see Monks and Minow 2001).

In case management promises not to exploit the firm's resources is in the terminology of game theory, simply not credible. Moreover, agency costs may be reduced if a blockholder exercises active ownership, which may benefit all the other minority shareholders. However, whether concentrated ownership is beneficial for all the shareholders or even society is still a theoretical and empirical unsettled question.

In the following concentrated ownership and investor protection are discussed which is followed by different ownership types, i.e., family and foundation ownership as well as managerial ownership. The entry ends with discussion of the increasing role of institutional investors, which also includes key empirical contributions.

Concentrated Ownership and Investor Protection

In the last decade there has been an increasing attention to the issue of investor protection, in particular how investor protection influences the development and functioning of capital markets globally. Shleifer and Vishny (1997) argue that concentrated ownership and investor protection may serve as important mechanisms to reduce agency costs due to the separation of ownership and control.

Law and economics is the study of how legal rules and court practice affect parties' incentives. The legal setup that regulates investor protection such as in company law, security regulation, and

accounting standards is enormous. However, the important issue here is that changes in the legal setup change the balance of power between management and investors, i.e., shareholders and creditors. If investor protection is increased, the development of financial market is increased as investors are more likely not to be exploited by top management. In relation to concentrated ownership, investor protection concerns how minority shareholders are formally protected.

In a study by La Porte et al. (1998), the authors relate legal families to the level of investor protection and ownership concentration. They find concentration of ownership of shares in the largest public companies is negatively associated with investor protection, consistent with the hypothesis that small diversified shareholders are unlikely to be important in countries that fail to protect their rights. Their cross-sectional regression model is estimated under the assumption that ownership depends on the rule of law, so that the rule of law is treated as an exogenous variable. However, it is not implausible that the causation may run in the opposite direction; hence both rule of law and ownership are considered as endogenous variables. The reason is that powerful owners might seek to influence the degree of investor protection by lobbying the political process. For instance, large shareholders with sufficient political power may try to evade the protection of minority shareholders stipulated in the company laws. It may be plausible that in countries with high ownership concentration, majority shareholders such as influential families would try to undermine investor protection through the political process or by influencing the court system.

Ownership by Families and Foundations

Family ownership plays a major role, not only in privately held firms but also in listed firms where families such as the founder or his decedents hold major ownership stakes in listed firms. Sometimes they maintain control holding shares with superior voting rights, whereas the shares with less voting power are held by other investors, e.g., institutional investors (see Rose 2008). Family ownership may facilitate long-term stable investors, but there is also a risk that the family may seek to enjoy private benefits at the expense of all the minority shareholders.

The relation between ownership and investor protection has also been analyzed in La Porta et al. (2002) who relate concentrated ownership with family ownership. They argue that resistance against the introduction of strong investor protection laws in some countries comes from families that control large corporations. They mention that "From the point of view of these families, an improvement in the rights of outside investors is first and foremost a reduction in the value of control due to the deterioration opportunities." Obviously, one cannot neglect the risk that, e.g., an incumbent family would seek control by occupying the board deriving private benefits of control. However, whether there exist severe agency costs associated with family control/control is in essence an empirical question. And, from this perspective, the recent evidence suggests the contrary. Anderson and Reeb (2003) conduct a substantial study of the firms in the S&P 500 where they show that family ownership is both prevalent and substantial, as family ownership is present in one third of the firms and accounts for 18% of outstanding equity. Specifically, they document that family firms perform better compared to nonfamily firms (also indication that the relationship is nonlinear). Moreover, the authors show that when family members serve as CEO, performance is improved compared to outside CEOs suggesting that family ownership may be an effective organization structure.

A new study by Isakov and Weisskopf (2014) also shows that family firms are more profitable than companies that are widely held or have a nonfamily blockholder. Their sample covers Swiss listed firms from 2003 to 2010, and the authors measure performance by Tobin's q. Specifically, they document that the generation of the family and the active involvement of the family play an important role for market valuation. Their sample covers Swiss firms and a country where investor protection is well developed. However, one cannot reject that in countries with

weaker investor protection, family ownership is not by per se positive for performance.

Foundation or trust ownership is widespread in several countries especially in Northern Europe, but it is also present in the Netherlands and Germany. Recent evidence does not associate foundation ownership with significant agency costs, even though the concept of foundation ownership seems to violate classical principal-agent theory (see, e.g., Thomsen and Rose 2004). A company founder may create a separate legal entity, i.e., a trust of a foundation, instead of transferring his shareholdings to his descendants. The foundation is governed by the board, but it has no owners, and the main purpose is to maintain control with the listed firm. The dividend proceedings are very often allocated to charitable purposes and/or the founder's descendants. The principal-agent model assumes that a principal hires an agent to conduct a task. Due to imperfect information, moral hazard problems may be created as the principal is not able to monitor the agent perfectly. Therefore, in order to align the interests between the principal, i.e., the owner and the agent, the latter is offered incentive contracts. However, in a foundation there are not any principals, since a foundation does not have any owners.

Managerial Ownership

Managerial control with the firm may be initiated by an MBO (management buyout). In such a situation the existing management acquires the shares which is financed by a high degree of debt. The literature concerning the impact of managerial ownership on firm performance does not offer a picture of unanimity.

Instead two divergence hypotheses exist. The so-called convergence of interest's hypothesis states a positive relationship between managerial ownership and firm performance. The underlying idea is to let a manager's remuneration depend more on the total wealth creation in the company by making him a residual claimant. As managers' stake rise, managers pay a larger share of the costs from activities that reduce firm value and therefore agency problems become less likely.

This stands in contrast to the *entrenchment hypothesis* that predicts a negative relationship between managerial ownership and firm performance. A manager who controls a substantial part of the firm's equity may be able to have sufficient influence to secure the most favorable employment conditions, including an attractive salary. One may argue that such benefits may also be obtained by his reputation, superior qualifications, or personality, independently of how much voting power a manager controls from his equity stake in the firm (see, e.g., Rose 2005).

Thus, even if managerial equity stake in a firm is low, there might be other forces to discipline managers away from opportunistic behavior such as competition in product markets, the managerial labor market. But at higher levels of managerial ownership, managerial entrenchment blocks takeovers making them more costly, which eventually decrease firm value since the probability of a successful tender offer decreases.

The Increasing Role of Institutional Investors

The increase in institutional funds in the industrialized countries has been tremendously rapid within the last 30 years, where institutional investors are the largest owners managing an enormous amount of capital. As a result, institutional investors are as a group considered the most influential actor on the scenes of capital markets. In the debate over corporate governance and, in particular, the role of institutional investors, more pressure has been put on institutional investors to be more active in their ownership, e.g., to exercise their proxies and their voting rights at the firm's general meeting.

Institutional investors cover a wide group of heterogeneous investors, which are all subjected to different legislations. They include pension funds, banks, insurance companies, mutual funds, mutual companies, and investment funds/foundations. Some of the largest institutional investors have all been very active in exercising their rights as owners, e.g., CalPERS, New York City pension fund, and TIAA-CREF. They have

sought, e.g., to challenge excessive executive compensation, the adoption of takeover defenses, to split the roles of chairman and CEO, and to ensure enough independent directors. However, on overall, institutional activism has been limited. The reason is that regulation often puts various restrictions on the ownership by institutional investors, such as requiring them not to have a dominant stockholding in given firm.

Moreover, when institutional investors hold a substantial proportion of shares, this might discipline management, since the free rider problem associated with dispersed ownership would be alleviated. Contrary to small investors, institutional investors are more able to absorb the costs from monitoring management and engaging in active ownership. Specifically, institutional investors may reduce the free rider problem caused by dispersed ownership and therefore avoid managerial focus on short termism. However, one may argue that if all small investors believe that institutional investors will undertake the monitoring role, the free rider problem may be enhanced. The reason is that this would destroy the incentives for small investors to play any active role at all.

Proponents of institutional activism argue that strengthening institutional investor ownership would benefit society as a whole, because they would be able to influence managerial actions, so that the interests of the society and the company more coincide; see, e.g., Monks and Minow (2001) for this view. Moreover, it has been advocated that institutional investors may facilitate the promise of "relationship investing" (see, e.g., Blair (1995)), who describe a situation where they are engaged in overseeing management over the long term instead of being detached or passive.

Furthermore, one may hypothesize institutional investors could influence management, only to take the interests of shareholders into account but also to serve the interests of other stakeholders; see Rose (2007). To illustrate, consider, e.g., a pension fund, where all the members have strong preferences against firms that directly or indirectly use child labor. Even if such firms may earn a higher profit due to lower costs, it might be reasonable for a pension fund not to invest in such firms, due to the members' strong preferences against the use of child labor. In other words, institutional investors may consider a broader view, trying to get management not only to care about the shareholders' interests, which to some extent can be justified, since shareholders are usually considered residual claimants (see, e.g., Fama and French 1983).

Moreover, one could argue that enhanced ownership by institutional investors does not necessarily influence performance positively. Specifically, it is doubtful that institutional investors act, as they have a long investment horizon. The reason is that a portfolio manager employed by institutional investors is evaluated yearly, by comparing each portfolio manager's performance, with a selected peer group or benchmark. As a consequence, the portfolio manager might care less about the return from their investments in the future 30 years from now, i.e., when the proceeds are repaid to the pension customers. Put differently, there is an embedded agency problem within institutional investors.

Furthermore, it is also questionable whether institutional investors would always act in a way that benefits all investor groups. Naturally, management in listed firms needs to care about the preferences of the large shareholders, since they are the owners and could replace incumbent management at the forthcoming general meeting. If institutional investors hold a high stake in a company, there is an inherent risk that institutional investors might seek to derive private benefits on behalf of all the other minority shareholders. For instance, institutional investors might get inside information, when management holds investor meetings or is in contact with the dominant owners. Even though this is prohibited by law, it is still quite difficult to prove afterwards by the authorities. Collusion between large blockholders and management may be sanctioned by the law, since this could violate the principle of the equal treatment of shareholders that prevails in most countries' legislation.

Some Key Empirical Contributions

Hartzell and Starks (2003) argue that institutional investors serve a monitoring role in mitigating the

agency problem between managers and shareholders. Specifically, they find that institutional ownership is positively related to the pay for performance sensitivity of executive compensation and negatively related to the level of compensations. The result is robust when they control for firm size, industry, and investment opportunities.

Duggal, R. and J. Millar (1999) empirically challenge the ability of institutional investors to monitor management. Based on takeover decisions in 1985–1990, they examine the impact of institutional ownership and performance, but they do not find evidence that active institutional investors, as a group, enhance efficiency in the market for corporate control. Duggal and Millar also identify a number of institutional investors that have a reputation for exercising an active ownership, but regressing bidder returns against active institutional investors only result in an insignificant relationship.

Wahal and McConnell (2000) find no support for the contention that institutional investors cause managers to behave myopically. Based on a large sample from 1988 to 1994 of US firms, they document a positive relation between the industry-adjusted capital expenditures, as well as research and development, and the proportion of shares held by institutional investors. Both are proxies for management's degree of long-term orientation.

Prevost and Rao (2000) study whether institutional investor activism benefits shareholders, using an event study of shareholder proposals surrounding proxy mailing dates. Contrary to earlier studies they find a strong negative wealth effect surrounding the proxy mailing dates of firms targeted by two very visible, publicity-seeking types of sponsors: CalPERS and coalitions of public funds sponsoring or cosponsoring one or more proposals on the same proxy. Prevost and Rao argue that the results are consistent with the hypothesis that a formal proposal submission signals a breakdown in the negotiation process between the funds and management.

Louis, Chan, and Lakonishok (1993) examine the price effect of institutional stock trading and they find that the average effect is small. They also document market asymmetry between price impact of buys versus sells, which is related to various hypotheses on the elasticity of demand for stocks, the costs of executing transactions, and the determinants of market impact. For instance, they argue that institutional purchase might be a stronger signal of favorable information, whereas there are many liquidity-motivated reasons to dispose a stock; see also Sias and Starks (1997) for an analysis of return autocorrelation and institutional investors.

Bhagat and Black and Blair (2004) conduct a large study of ownership and performance over a 13-year period, focusing on whether relationship investing has a positive impact on firm performance. They document a significant secular increase in large-block shareholding with sharp percentage increase in these holdings by mutual funds, partnerships, investment advisors, and employee pension plans. However, most institutional investors, when they purchase large blocks, sell the blocks relatively quickly afterwards. Bhagat, Black, and Blair provide a mixed result of whether relational investing affects firm performance. In the late 1980s where there was a high takeover wave, there is a significant relation between relational investing and firm performance, but this pattern was not found in the other periods. In essence, they do not find any persistent and sustainable effect of relational investing on firm performance. Thus, they argue that the idea of relational investing must be more carefully specified in theory.

Ackert and Athanassakos (2001) focus on agency considerations within institutional investors. They show that market frictions are important concerns for institutional investors, when they make portfolio allocation decisions. The availability of information about a firm is a significant friction, so that institutional holding increases with market value and firm's visibility, as proxied by the number of analysts following the firm. They also show that institutions adjust their portfolios away from highly visible firms at the beginning of the year but increase their holdings in these firms as the year-end approaches, which is as they argue consistent with the gamesmanship hypothesis.

In contrast to several other studies that focus on firm-level effects of institutional ownership, Davis (2002) examines how institutional shareholding in the largest countries on aggregate level impacts macroeconomies. Specifically, Davis links the development of institutional investors to important indicators of corporate sector performance, such as increasing dividend distribution, less fixed investment, and higher productivity growth.

Anand et al. (2013) examine the impact of institutional trading on stock resiliency during the financial crisis of 2007–2009. A resilient market is defined as one where prices recover quickly after a liquidity shock. The authors focus on why financial markets stayed illiquid over an extended period during the 2007–2009 crisis. They show that liquidity suppliers withdraw from risky securities during the crisis, and their participation does not recover for an extended period of time. Moreover, the illiquidity of specific stocks is significantly affected by institutional trading patterns.

Institutional shareholders exercise their formal power by voting at the AGM. The agenda on the AGM consists mostly by the board's own proposals. However, there has been an increasing tendency of shareholder-initiated proxy proposals. Renneboog and Szilagyi (2011) study the role of shareholder proposals in corporate governance. They find that target firms tend to underperform and have generally poor governance structures with little indication of systematic agenda setting by the proposal sponsors. The authors also find that proposal implementation is largely a function of voting success but is affected by managerial entrenchment and rent seeking. According to the authors, their results imply that shareholder proposals are a useful device of external control which should not be legally restricted.

References

Ackert LF, Athanassakos G (2001) Visibility, institutional preferences and agency considerations. J Psychol Financ Markets 2:201–209

Anand A, Irvine P, Puckett A, Venkataraman K (2013) Institutional trading and stock resiliency: evidence from the 2007–2009 financial crisis. J Financ Econ 108:773–793

Anderson RC, Reeb DM (2003) Founding-family ownership and firm performance. Evidence from the S&P 500. J Financ LVIII(3):1301–1328

Bhagat S, Black B, Blair M (2004) Relational investing and firm performance. J Financ Res 27:1–30

Blair M (1995) Ownership and control. Rethinking corporate governance for the twenty first century. The Bookings Institution, Washington, DC

Davis EP (2002) Institutional investors, corporate governance and the performance of the corporate sector. Econ Sys 26:202–229

Duggal R, Millar JA (1999) Institutional ownership and firm performance: the case of bidder returns. J Corp Financ 5:103–117

Fama E, French K (1983) Agency problems and residual claims. J Law Econ 26:327–349

Hartzell JC, Starks LT (2003) Institutional investors and executive compensation. J Financ 58: 2351–2374

Isakov D, Wisskopf J-P (2014) Are founding families special blockholders? An investigation of controlling shareholder influence on firm performance. J Bank Financ 41:1–16

La Porta R, de-Silanes L, Schleifer A, Vishny RW (1997) Legal determinants of external finance. J Financ 52:1131–1150

Louis K, Chan C, Lokonishok J (1993) Institutional trades and intraday stock price behavior. J Financ Econ 33:173–199

Monks RAG, Minow N (2001) Corporate governance, 2nd edn. Blackwell Business, Malden

Porta L, Rafael FL-d-S, Shleifer A, Vishny RW (1998) Law and finance. J Polit Econ 61: 1113–1155

Porta L, Rafael FL-d-S, Shleifer A, Vishny RW (2002) Investor protection and corporate valuation. J Financ 57:1147

Prevost AK, Roa RP (2000) Of what value are shareholder proposals sponsored by public pension funds? J Bus 73:177–204

Renneboog L, Szilagyi PG (2011) The role of shareholder proposals in corporate governance. J Financ Econ 17:167–188

Rose C (2005) Managerial ownership and firm performance of listed Danish firms – in search of the missing link. Eur Manag J 23(5):542–553

Rose C (2007) Can institutional investors fix the corporate governance problem? – Some Danish evidence. J Manag Gov 11:405–428

Rose C (2008) A critical analysis of the one share –- one vote controversy. Int J Disclos Gov 5(2): 126–139

Shleifer A, Vishny RW (1997) A survey of corporate governance. J Financ 52:737–783

Thomsen S, Rose C (2004) Do companies need owners? Foundation ownership and firm performance. Eur J Law Econ 18:343–363

Wahal S, McConell JJ (2000) Do institutional investors exacerbate managerial myopia? J Corp Financ 6: 307–329

Confiscation Orders and Judicial Cooperation in the EU

Barbara Piattoli Girard
Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, Università degli Studi del Piemonte Orientale "Amedeo Avogadro", Alessandria, Italy

Abstract

Following ongoing reflection and experience at European level, it is possible and necessary to reason about the main features of the EU's legal strategy in building a simplified procedure for the recognition of confiscation orders among EU countries, in order to avoid the different barriers to the effectiveness of the EU's regime on the confiscation of proceeds of crime. It's significant in this context to focus on the consequences of the principle of mutual recognition on the rights of individuals. The proposal for a new regulation on the mutual recognition of freezing and confiscation orders aims to amend the EU's regime and eliminate gaps, uncertainties that legal rules still present; however, its adoption might significantly improve effectiveness of the EU's action if the emphasis on legal solutions doesn't come at the expense of broader questions concerning the safeguards applicable to domestic criminal proceedings which are crucial to ensuring effective cooperation between Member States in recovery action.

EU Legal Strategy on Mutual Recognition of Confiscation Orders: Improving an Effective Cooperation Between Member States?

The confiscation and recovery of criminal assets have assumed a prominent position in the fight against organized and other profit-driven crime in the EU, and it is also an important tool to combat terrorist financing (COM(2016)50 final, Chapter 1.3). The terrorist attacks in 2015 and 2016 in the European Union and beyond underlined the urgent need to prevent and fight terrorism through a rapid and cohesive action to disrupt terrorist financing and its close link with organized crime networks. This has reinforced the issue of a better coordination and cooperation at EU level in order to give effectiveness of the EU's legal strategy in building a genuine area of justice, which led to the adoption of several measures designed to improve confiscation across the EU Member States with the most recent instruments of mutual recognition (Lelieur 2015).

Through the Council Framework Decision 2006/783/JHA of 6 October 2006 concerning the application of the principle of mutual recognition of confiscation orders (together with the Framework Decision 2003/577/JHA of 22 July 2003 on the execution in the European Union of orders freezing property or evidence), the EU is intended to strengthen the direct execution of confiscation (and freezing) orders for proceeds of crime (Brown 1996) by establishing simplified procedures for recognition among EU countries and rules for dividing confiscated property between the country issuing the confiscation order and the one executing it. In addition the recent proposal for a regulation of the European Parliament and of the Council on the mutual recognition of freezing and confiscation orders (COM(2016)819 final), which will replace these two framework decisions, builds an existing EU legislation on mutual recognition considering the need of putting in place a comprehensive system for freezing and confiscation of proceeds and instrumentalities of crime in the EU (Kingah 2015). It addresses the fact that legislation on mutual recognition of freezing and confiscation orders leaves *lacunae,* and Member States have developed new forms of freezing and confiscation of criminal assets (Fazekas and Nanopoulos 2016).

The European Council, meeting in Tampere on 15 and 16 October 1999, stressed that the principle of mutual recognition should become the cornerstone of judicial cooperation in both civil and criminal matters within the union (Flore 2014; Vermeulen 2014). After the entry into force of the Lisbon Treaty, confiscation has been given strategic priority at EU level as an effective instrument to fight organized crime (Feraldo Cabana

2014). The purpose of the actual European rules is to facilitate cooperation between Member States as regards the mutual recognition and execution of orders to confiscate property so as to oblige a Member State to recognize and execute in its territory confiscation orders issued by a court competent in criminal matters of another Member State.

Mutual recognition instruments are linked to harmonization measures (Council Framework Decision 2005/212/JHA of 24 February 2005 on confiscation of crime-related proceeds, instrumentalities, and property and Directive 2014/42/EU of 3 April 2014 on the freezing and confiscation of instrumentalities and proceeds of crime in the EU) (Simonato 2015). Both types of instruments (harmonization measures and mutual recognition instruments) are necessary in order to have a functioning regime of recovery of criminal assets, and they complement each other. In parallel, efforts were made to strengthen the identification and tracing of the proceeds and instrumentalities of crime. Council Decision 2007/845/JHA provides for the establishment of Asset Recovery Offices in all Member States.

In particular, the Directive 2014/42/EU, which replaces certain provisions of Council Framework Decision 2005/212/JHA, provides uniform rules for domestic possibilities to freeze and confiscate assets. Whereas the Framework Decision 2005/212/JHA continues to apply to all criminal offenses punishable by detention of at least 1 year, regarding ordinary confiscation, the Directive could only cover the so-called Eurocrimes and set minimum rules for national freezing and confiscation regimes: it requires ordinary and value confiscation for Eurocrimes, including where the conviction results from proceedings in absentia. It provides rules for extended confiscation subject to certain conditions (Boucht 2013). It also enables confiscation where a conviction is not possible because the suspect or accused person is ill or has absconded (Alagna 2015). The Directive also enables the confiscation of assets in the possession of third parties. Finally, the Directive introduces a number of procedural safeguards, such as the right to be informed of the execution of the freezing order including, at least briefly, on the reason or reasons, the effective possibility to challenge the freezing order before a court, the right of access to a lawyer throughout the confiscation proceedings, the effective possibility to claim title of ownership or other property rights, and the right to be informed of the reasons for a confiscation order and to challenge it before a court.

Playing field of the Framework Decision 2006/783/JHA is mutual recognition and direct execution of confiscation orders. It acknowledges the need to modernize cross-border cooperation legislation within the EU and at an international level providing Member States with a set of procedural rules for a coordinate action (for the execution in the European Union of orders freezing property or evidence, see Council Framework Decision 2003/577/JHA of 22 July 2003. Both instruments are based on the principle of mutual recognition and work in a similar way). The Framework Decision 2006/783/JHA requires confiscation orders issued in one Member State to be recognized and executed in another Member State. The orders are transmitted alongside a certificate to the competent authorities in the executing State which must recognize them without further formalities and take the measures necessary for their execution.

Like it happens for the European arrest warrant, in order to facilitate the procedure, the double criminality check has been abolished in relation to a list of 32 offenses, provided that these offenses are punishable in the issuing country by a custodial sentence of at least 3 years. For all types of crime other than those listed in the Framework Decision 2006/783/JHA, the executing country can continue to apply the principle of double criminality – that is, it can make recognition and execution of the order dependent on the condition that the facts giving rise to the confiscation order constitute an offense according to its law. In limited cases the executing country may refuse to recognize and execute the order if the certificate is missing or incomplete or does not correspond to the order (in accordance with the *ne bis in idem* principle, the same person has already been the subject of a confiscation order for the same facts, for immunities or privileges that prevent execution, and for the rights of the parties

concerned and third persons acting in good faith); if the judgment was given in the absence of the person concerned, unless he/she was informed of the date and place of the trial and that an order may be handed down regardless of his/her presence; if he/she was represented by a legal counselor; or if he/she did not contest the judgment nor request a retrial or an appeal within the set time limit, for the principle of territoriality.

According to the procedural rules, the confiscation order, together with a certificate of which a copy is annexed to the framework decision and must be translated into the official language of the executing country, or another official language of the EU as indicated by that country, will be sent directly to the competent authority of the EU country where the natural or legal person concerned has property or income, is normally a resident, or has its registered seat. If the issuing authority cannot identify the authority in the executing country that is competent to recognize and execute the order, it will make inquiries, including through the European Judicial Network. A written record of the transmission of the order must be available to the executing country, which checks that it is genuine.

The transmission of a confiscation order does not restrict the right of the issuing country to execute the order itself. Where appropriate, the competent authority in the executing country must be informed.

The executing country recognizes and executes the order forthwith and without requiring the completion of any further formalities. The order is executed in accordance with the law of the executing country and in a manner decided upon by its authorities. The amounts confiscated are disposed of by the executing country as follows: if the amount is below EUR 10000, it accrues to the executing country; if it is above that amount, 50% of it is transferred to the issuing country. Both the executing and the issuing country can grant a pardon or amnesty, while the issuing country alone is responsible for appeals on the substance of the case lodged against the order.

This EU's intervention based on the mutual recognition aims to give more effectiveness of the EU's asset confiscation regime. The experience has shown that not all the Member States have transposed the framework decisions on mutual recognition of freezing and confiscation orders until now, and the level of transposition of these framework decisions into the national legal systems of Member States was not satisfactory. It has been complained about the underuse of confiscation in cross-border situations at judicial level with the current system (commission staff working paper accompanying document to the proposal for a Directive of the European Parliament and the Council of the freezing and confiscation for proceeds of crime in the European Union, SWD(2012) 31 final; Report from the Commission to the European Parliament and the Council of 22.12.2008 [COM(2008) 885 final; Report from the Commission to the European Parliament and the Council of 23.8.2010 [COM(2010) 428 final].

It appears the need of a more strategic action at EU level to improve mutual recognition of freezing and confiscation orders which is also relevant to address terrorist financing in a more effective manner. The European initiative COM(2016)819 of 21 December 2016 is a response to these calls, and it addresses the fact that Member States have developed new forms of freezing and confiscation of criminal assets. It also takes into account developments at EU level, including the minimum standards set out in Directive 2014/42/EU. Whereas the directive improves the domestic possibilities to freeze and confiscate assets, the proposal aims to improve the cross-border enforcement of freezing and confiscation orders. Together, both instruments should contribute to effective asset recovery in the European Union.

Opportunity and Possible Benefits of the New Proposal for an EU Regulation on Mutual Recognition of Freezing and Confiscation Orders

The proposed regulation, once adopted, will be directly applicable in the Member States. This provides greater legal certainty and avoids the transposition problems that the framework decisions on mutual recognition of freezing and

confiscation orders were subject to. A regulation does not leave a margin to Member States to transpose such rules with the result that orders issued by other Member States will have to be executed like domestic ones, without the need to modify their internal legal system and their way of working.

Another important issue is the extended scope of the instrument compared to Directive 2014/42/EU as the proposal improves provisions that ensure a wider circulation of freezing and confiscation orders imposed by a court following proceedings in relation to a criminal offense including non-conviction-based confiscation orders. So the regulation applies to all types of orders covered by Directive 2014/42/EU (including extended confiscation and third-party confiscation orders), and in addition, it will also cover orders for non-conviction-based confiscation issued within the framework of criminal proceedings: the cases of death of a person, immunity, prescription, and cases where the perpetrator of an offense cannot be identified, or other cases where a criminal court can confiscate an asset without conviction when the court has decided that such asset is the proceeds of crime. This requires the court to establish that an advantage was derived from a criminal offense. In order to be included in the scope of the regulation, these types of confiscation orders must be issued within the framework of criminal proceedings. Therefore all safeguards applicable to such proceedings will have to be fulfilled in the issuing State. This perspective could generate several problems as sometimes orders can be issued without all the safeguards applicable to criminal proceedings (European Court of Human Rights, 22 October 2009, Paraponiaris c. Greece, n. 42,132/06; see, *infra*, § 3). However, the regulation does not cover civil and administrative orders. For what concerns its object area, it is not limited to particularly serious crime with a cross-border dimension so-called "Eurocrimes" (unlike Directive 2014/42/EU which is based on 83 TFEU) as Article 82 TFEU (on which this proposal is based on) does not require such a limitation for mutual recognition of judgments in criminal matters.

Fundamental Rights

Asset recovery is assumed to have positive impacts to enhance redistributive justice for victims of crime, and the victim's right to compensation and restitution has been duly taken into account in the proposal. It is ensured that in cases where the issuing State confiscates property, the victim's right to compensation and restitution has priority over the executing and issuing States' interest.

This issue underlined that freezing and confiscation measures may interfere with fundamental rights protected by the EU Charter of Fundamental Rights (the Charter) and the European Convention on Human Rights (ECHR). In particular, the European Court of Human Rights (ECtHR) has repeatedly considered non-conviction-based confiscation, including civil and administrative forms, and extended confiscation to be consistent with Article 6 ECHR and Article 1 of Protocol 1, if effective procedural safeguards are respected. Shifts of the burden of proof concerning the legitimacy of assets have not been found in violation of fundamental rights by the ECtHR, as long as they were applied in the particular case with adequate safeguards in place to allow the affected person to challenge these rebuttable presumptions (ECtHR 29 ottobre 2013, Varvara c. Italia; ECtHR 20 gennaio 2009, Sud Fondi et al. c. Italy; ECtHR 10 maggio 2012, Sud Fondi e altri c. Italy). Recently, ECtHR 23 February 2017, De Tommaso c. Italia focused on preventive measures, and the European Court reiterates its settled case law, according to which the expression "in accordance with law" not only requires that the impugned measure should have some basis in domestic law but also refers to the quality of the law in question, requiring that it should be accessible to the persons concerned and foreseeable as to its effects. "One of the requirements flowing from the expression 'in accordance with law' is foreseeability. Thus, a norm cannot be regarded as a 'law' unless it is formulated with sufficient precision to enable citizens to regulate their conduct; they must be able – if need be with appropriate advice – to foresee, to a degree that is reasonable in the circumstances, the

consequences which a given action may entail." The Court reiterates that a rule is "foreseeable" when it affords a measure of protection against arbitrary interferences by the public authorities. A law which confers a discretion must indicate the scope of that discretion, although the detailed procedures and conditions to be observed do not necessarily have to be incorporated in rules of substantive law (see Silver and Others v. the United Kingdom, 25 March 1983, § 88, Series A no. 61).

This statement can also involve the determination of preventive measure concerning real property ("in rem"); they apply – if requested by the prosecutor and ordered by the judge – before, the conviction becomes final under special circumstances, and one of the requirements flowing from the expression "in accordance with law" is foreseeability. According to this perspective, the domestic law cannot be vague and excessively broad terms. The individuals to whom preventive measures are applicable and the content of these measures must be defined by law with sufficient precision and clarity.

Anyway, some important safeguards are included in the proposed regulation of the European Parliament and of the Council: the principle of proportionality needs to be respected; there are grounds for refusal based on the non-respect of the principle of "ne bis in idem" and the rules on "in absentia" proceedings. Moreover, the rights of bona fide third parties have to be respected; there is an obligation to inform interested parties of the execution of a freezing order including the reasons thereof and the legal remedies available, and there is an obligation for Member States to provide for legal remedies in the executing State. Furthermore, Article 8 of Directive 2014/42/EU includes a list of safeguards that need to be ensured by the Member States for those orders falling within the scope of the directive.

Finally, all criminal law procedural safeguards are applicable. This includes in particular the right to a fair trial enshrined in Article 6 ECHR and Articles 47 and 48 of the Charter. It also includes the relevant legislation at EU level on procedural rights in criminal proceedings:

- Directive 2010/64/EU on the right to interpretation and translation in criminal proceedings
- Directive 2012/13/EU on the right to information about rights and charges and access to the case file
- Directive 2013/48/EU on the right of access to a lawyer and communication with relatives when arrested and detained
- Directive 2016/343 on the strengthening of certain aspects of the presumption of innocence and the right to be present at one's trial
- Directive 2016/800 on the procedural safeguards for children
- Directive 2016/1919 on legal aid for suspects and accused persons in criminal proceedings and for requested persons in European arrest warrant proceedings

However, Member States often have regulation which does not assure a coherent exercise of the right of defense, in particular if the confiscation order is issued in preliminary steps of the proceedings such as a summary judgment or a dismissal of the case (e.g., or an acquittal anticipated *ex* art. 129 c.p.p. and 469 c.p.p. in the Italian procedural system). That lack of safeguards could be exploited for a refusal in certain procedural situations of a cohesive cooperation system for freezing and confiscation of proceeds and instrumentalities of crime.

The New Rules for the Procedure of Mutual Recognition

The procedure for recognition and execution of freezing and confiscation orders is regulated separately in the proposal to simplify direct application by competent national authorities. As it regards in particular confiscation orders, it provides for a direct transmission of a confiscation order between competent national authorities but also allows for the possibility of assistance by central authorities. The confiscation order must be accompanied by a standard certificate annexed to this proposal. The executing authority must recognize the confiscation order without further formalities and must take the necessary measures

for its execution in the same way as for a confiscation order made by an authority of the executing State unless it invokes one of the grounds for refusal or postponement. A list of offenses for which the mutual recognition and execution of freezing and confiscation orders cannot be refused based on dual criminality is the same as the list contained in other mutual recognition instruments with one exception only: point (y) of the list now reflects the existence of common minimum standards for combating fraud and counterfeiting of noncash means of payment (Framework Decision 2001/413/JHA). Dual criminality cannot be invoked for a list of offenses punishable by at least 3 years of imprisonment in the issuing State. In cases of offenses not included in this list, recognition can be refused if the crime to which the freezing or confiscation order relates is not a criminal offense under the laws of the executing State.

A list of grounds for nonrecognition and nonexecution of confiscation orders on which basis the executing authority may refuse the recognition and execution of the confiscation order is laid down in Article 9. The list differs significantly from the list contained in the 2006 Framework Decision. Some grounds for refusal remain the same, e.g., the ground based on the principle *ne bis in idem* or the ground based on immunity or privilege. However, the grounds for refusal linked to the type of the confiscation order (e.g., extended confiscation) have not been included in the proposal thus considerably broadening and strengthening the mutual recognition framework.

Regarding the ground for refusal based on the right to be present at the trial, it only applies to trials resulting in confiscation orders linked to a final conviction and not to proceedings resulting in non-conviction-based confiscation orders. Detailed rules for a possibility to confiscate different type of property than the one specified in the confiscation order are laid down. The instrument lays down conditions for issuing and transmitting a freezing order to align the proposal with Article 6 of Directive 2014/41/EU, thereby ensuring that the same conditions apply to freezing for evidence and freezing for subsequent confiscation.

Conclusion

The proposal for a new regulation aims to amend the EU regime and to eliminate gaps, uncertainties that legal rules still present; however, its adoption might significantly improve effectiveness of the EU's action if the emphasis on legal solutions does not come at the expense of broader questions concerning the safeguards applicable to criminal proceedings which are crucial to ensuring effective cooperation between Member States in recovery action. It is significant in this context to focus on the consequences of the principle of mutual recognition on the rights of individuals and to consider the need, in particular in relation to non-conviction-based confiscation, of effective procedural safeguards within all the proceedings steps according to the ECtHR case law. This perspective will oblige Member States to rethink and to restyle certain domestic procedural rules to give real effectiveness of the EU's action, even if the proposed regulation, once adopted, will be directly applicable in the Member States.

Reference

Alagna F (2015) Non-conviction based confiscation: why the EU directive is a missed opportunity. European Journal on Criminal Policy and Research 21(4):447

Boucht J (2013) Extended confiscation and the proposed directive on freezing and confiscation of criminal proceeds in the EU: on striking a balance between efficiency, fairness and legal certainty. European Journal of Crime Criminal Law and Criminal Justice 21(2):127

Brown AN (1996) Proceeds of crime, money laundering, confiscation and forfeiture. W. Green/Sweet&Maxwell, Edimburgh

Fazekas M, Nanopoulos E (2016) The effectiveness of EU law: insights from the EU legal framework on asset confiscation. European Journal of Crime, Criminal Law and Criminal Justice 24(1):39

Feraldo Cabana P (2014) Improving the recovery of assets resulting from organised crime. European Journal of Crime, Criminal Law and Criminal Justice 22:13

Flore D (2014) Droit pénal européen: les enjeux d'une justice pénale européenne. Larcier

Kingah S (2015) Measures for asset recovery: a multiactor global fund for recovered stolen assets. World Bank Legal Review 6:457

Lelieur J (2015) Freezing and confiscating criminal assets. European Union in Criminal Law Review 5(3):279

Simonato M (2015) Directive 2014/42/EU and social reuse of confiscated assets in the EU: advancing a culture of legality. New Journal of European Criminal Law 2:195
Vermeulen G (2014) Essential texts on international and European criminal law. Maklu, Antwerp-Apeldoorn

Conflict of Interest

Remus Valsan
Edinburgh Law School, University of Edinburgh, Edinburgh, Scotland, UK

Definition

Fundamental to the notion of conflict of interest is the idea that someone's ability to exercise proper judgment is at risk of being affected by a personal interest or by a competing duty. These extraneous factors interfere with judgment not as ends that a decision-maker has in view but as factors that tend to influence the ends in view. The presence of such factors puts at risk the decision-maker's ability to evaluate the weight to be given to the relevant considerations on which the decision is based, irrespective of his desire to resist the temptation of self-interest. The main danger in a conflict of interest situation is the risk of unreliable judgment rather than corruption.

Introduction

A conflict of interest situation arises when a person who has a preexisting obligation to exercise judgment over the interests of another has a personal interest or duty that tends to interfere with the proper exercise of his judgment (Davis 1982). Conflicts of interest can arise in various situations where a preexisting duty of proper judgment exists. The legal relations that can create a situation of conflict of interest are not limited to positions with respect to which there are established rules against conflicts, such as per se fiduciaries, members of professions, or public officials (Norman and MacDonald 2010). The rules against conflicts of interest applicable to these roles are more visible because, on the one hand, such roles involve exercise of professional judgment or official discretion and, on the other hand, the maintenance of a good public image of such office holders is essential (Valsan 2016).

A situation of conflict of interest should be distinguished from a situation of conflicting interests. The former is a conflict between interest and judgment duty, whereas the latter is an opposition between the individual interests of parties to a legal relation. The two categories have distinct legal regimes. In private law, conflict of interest situations are governed by the law of fiduciary duties. Conflicting interest situations are pervasive, but the law leaves it largely to the parties to adjust their idiosyncratic interests through bargaining. The law protects the effectiveness of the bargaining process through several doctrines, such as good faith, undue influence, unconscionability, or disclosure obligations. The fundamental aims of the legal doctrines governing the two types of conflict are, thus, significantly different. While in a conflict of interest situation the law aims to protect the unencumbered judgment of the decision-maker, in a conflicting interest situation, the law aims to establish a certain level playing field between contracting parties.

Actual, Potential and Apparent Conflicts of Interest

A conflict of interest is actual if the decision-maker has a conflicting interest or duty with respect to a certain judgment that he must make. A conflict of interest is potential if the decision-maker has a conflict of interest with respect to a certain judgment but is not yet required to make that particular judgment (Davis 1998) or if a real sensible possibility of conflict exists (Valsan 2016).

Actual or potential conflicts of interest should be distinguished from situations that only give the appearance of a conflict of interest. A conflict of interest is apparent if there is no actual or potential conflict, but a third party may erroneously believe that a conflict exists. Appearances of conflict cannot, by themselves, indicate the existence of

a conflict. The outward impressions or indications that a person's actions produce are often a matter of the beholder's subjective perception (Davis 1998).

The distinction between apparent conflicts, on the one hand, and actual and potential conflicts of interest, on the other, is important as concerns the actions that decision-makers must take when faced with these situations. Apparent conflicts, although posing no actual or potential threat to the decision-maker's judgment, should nevertheless be clarified, for the same reason for which any apparent wrongdoing is objectionable. If the decision-maker becomes aware of appearances of conflict of interest relating to his activity, he must eliminate them by making available enough information to show that there is no actual or potential conflict. In the case of professionals or public officials, the obligation to dissipate appearances of conflict is justified by the damage that such appearances cause to public confidence in the profession as a whole (Davis 2001).

The Risk of Impaired Judgment

The legal rules governing conflicts of interest aim to eliminate the risk of unreliable judgment. In certain cases, such as public officials or members of a profession, the rules serve the added purpose of maintaining the appearance of propriety of judgment, in order to preserve the public confidence in the respective offices or professions (Valsan 2016).

The concept of judgment or discretion denotes the absence of a predefined algorithm based on which a decision can be modelled. In a situation requiring the exercise of judgment, the specification of the problem to be solved or the ends to be achieved are contested or open to interpretation. In contrast, decisions that do not require judgment are routine, mechanical, or ministerial. They require only technical rationality, in the sense of applying specific techniques or theories to achieve predefined unambiguous goals. Given the absence of a predefined pattern regarding the ends to be attained and the means to achieve them, the exercise of judgment goes beyond mechanical rule-following. When a decision requires judgment, different decision-makers may disagree on the ends to be pursued or on the optimal course of action, without anyone being wrong in an objective, measurable sense (Davis 1993, 2001).

When a person has an obligation to exercise judgment over the interests of another, the risk of an impaired decision-making process is not only pervasive, but also difficult to detect or correct. The literature on cognitive and motivational biases shows that the influence of an extraneous interest over the decision process is seldom a matter of deliberate choice. Information processing relating to self-interest is relatively effortless and unconscious, whereas the processes governing professional judgment are more analytical and more effortful. When professional judgment and self-interest point in opposite directions, self-interest often prevails, even when decision-makers consciously attempt to comply with their professional requirements (Moore and Loewenstein 2004). Since self-interest considerations are governed by automatic processes that tend to occur outside of conscious awareness, conflict of interest situations pose a problem even when the decision-maker does not exploit them in corrupt ways. Conflicting personal or professional interests can impair the judgment of even the most conscientious and devoted expert by influencing the way in which a decision-maker evaluates the seriousness of various risks, the desirability of certain outcomes, or the perception of connections between cause and effect (Norman and MacDonald 2010). Consequently, conflicts of interest are reprehensible not so much because they create a measurable bias but because they create an unusual risk of error of judgment, which cannot be adequately measured and corrected (Davis 1998).

Since perturbing interests affect the decision-making process as factors that tend to influence the ends in view, the extent of the effect of such interests on one's judgment cannot be assessed based on the actual decision taken. The decision-maker has discretion, in these sense of authority to decide the appropriate course of action, so one cannot simply measure the influence of the perturbing interest by comparing the decision adopted with an objectively right decision. Since the effect of a conflict of interest cannot be assessed based on results, the legal rules governing conflict of interest focus on certain kinds of identifiable interests that are particularly

threatening to the exercise of judgment, such as material interests or family ties. The categories of interfering interests, however, are not closed (Davis 2001).

Not all factors that might compromise one's judgment can be regarded as interfering interests. First, factors that affect a decision-makers' competence, such as the depth and accuracy of information used to adopt a decision, may be more relevant to the duty to exercise due skill and care than to conflicts of interest. Second, exercise of discretion allows, by definition, a certain degree of subjectivity in the decision-making process. The combination of personal characteristics that is specific for each decision-maker accounts for the diversity of equally valid results that can occur in a situation involving discretion. The line between legitimate factors that influence the decider's judgment and factors that have the ability to create a conflict of interest is sometimes blurry. Ultimately, what constitutes a conflict of interest in a particular situation is an empirical question (Stark 2000).

Managing Conflicts of Interest

People have an imperfect understanding of the effect of self-interest on their judgment and of the optimal way of correcting this influence. Individuals have little insight into their cognitive processes and may thus be unable to detect if, and to what extent, their judgment reflects self-interested motivations (Nisbett and Wilson 1977).

People tend to either underestimate or overestimate the biasing effect of self-interest on themselves. Because they cannot have an objective understanding of the effect of self-interest on their decision-making, people may tend to think that they can resist the effects of self-interest without their judgment being affected. They may be inclined to discount self-interest as their own motivation and overestimate the role of self-interest in motivating other people (Pronin 2006). The opposite is also possible. When people are aware of a situation where self-interest could plausibly intrude on their judgment, they may assume that the bias exists and is influencing them. In such cases, the more committed a decision-maker is to fairness and objectivity, the more likely he is to over-compensate for the presumed bias of self-interest, thus undermining the reliability of his decision (Kahneman et al. 1982).

Given the inability to measure accurately the effects of self-interest on judgment, the need to devise an effective response to a conflict of interest situation arises. Avoidance of conflicts is one potential solution. Persons having a duty to exercise judgment in the interest of another must avoid situations in which their interests pose an actual or potential threat to the reliability of their judgment. Although avoidance of conflict situations is an important duty of decision-makers, a flat prescription to avoid all conflicts of interest is undesirable. On the one hand, not all conflicts of interest are avoidable. Some conflict situations are embedded in the relation, while others arise independently of decision-maker's will. On the other hand, the mere fact of being in a situation of conflict is not always wrong from a legal or ethical perspective. Failure to address the conflict situation, however, is often reprehensible (Davis 1982).

Another potential strategy is to disclose the conflict to those relying on the conflicted person's judgment. Common sense suggests that complete disclosure will give the beneficiary of disclosure the opportunity to give informed consent to the situation of conflict, to adjust reliance accordingly, or to replace the decision-maker. Except the latter scenario, disclosure is an effective response if it does not affect the decision-maker's judgment process, or, alternatively, if the beneficiary is able correctly to adjust to the risk of impaired judgment. Psychological research shows that neither of these conditions may be met. Sometimes both parties may be worse off following disclosure (Cain et al. 2005).

Disclosure may have the unintended consequence of liberating a decision-maker from concerns about ethicality and give him a moral license to incorporate the conflicting interest into the decision-making process. Moreover, knowing that the beneficiary is likely to discount the decision to correct for the self-interest, the decision-maker may be tempted to counteract this adjustment by allowing self-interest to influence their decision even further. In addition, beneficiaries of disclosure do not always adjust to counteract the self-interest. In some cases, they see disclosure as

a sign of the decision-maker's trustworthiness and may increase their confidence in the latter's judgment (Cain et al. 2005).

Another potential response to the problem posed by a conflict of interest is to escape the situation. The decision-maker can escape a conflict by redefining the scope of the relationship, so that the scope of the judgment is restricted, by divesting himself of the interest creating the conflict or, where possible, by withdrawal from the relationship (Davis 1982). Divestment of the conflicting interest, however, may not be entirely effective. A person who in the past had an interest in the outcome of a decision cannot be said to be in the same psychological position as someone who was never interested in that matter (Miller 2005).

All potential responses to a situation of conflict of interest have their specific costs and benefits, and the law does prescribe an optimal response. In devising a strategy to manage conflicts, two guiding principles should be observed. First, a mere instruction to abstain from the temptation of self-interest is a mistaken response to a conflict of interest. Second, disclosure and consent do not put the parties in the same situation as when no conflict exists (Valsan 2016).

Conclusion

Personal interests or duties can affect the judgment of even the most honorable and disciplined decision-maker. A person has a conflict of interest on the basis of being in a conflicted situation, irrespective of that person's belief that he is capable of resisting the temptation or corrupting influence of the interest that could interfere with his judgment. Consequently, prescribing ethical self-restraint is a misguided solution to the conflict. Decision-makers must take active steps to steer clear of situations of conflict, to manage unavoidable ones, and to dissipate the mere appearances of conflict.

Cross-References

► Fiduciary Duties

References

Cain DM et al (2005) The dirt on coming clean: perverse effects of disclosing conflicts of interest. J Leg Stud 34:1
Davis M (1982) Conflict of interest. Bus Prof Ethics J 1:17
Davis M (1993) Conflict of interest revisited. Bus Prof Ethics J 12:21
Davis M (1998) Conflict of interest. In: Chadwick R (ed) Encyclopedia of applied ethics, vol 1. Academic Press, London, p 589
Davis M (2001) Introduction. In: Davis M, Stark A (eds) Conflict of interest in the professions. Oxford University Press, Oxford, p 3
Kahneman D et al (eds) (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge
Miller DT (2005) Psychologically naïve assumptions about the perils of self-interest. In: Moore DA et al (eds) Conflicts of interest: challenges and solutions in business, law, medicine, and public policy. Cambridge University Press, New York, p 126
Moore DA, Loewenstein G (2004) Self-interest, automaticity, and the psychology of conflict of interest. Soc Justice Res 17:189
Nisbett RE, Wilson TD (1977) Telling more than we can know: verbal reports on mental processes. Psychol Rev 84:231
Norman W, MacDonald C (2010) Conflicts of interest. In: Brenkert G, Beauchamp T (eds) The Oxford handbook of business ethics. Oxford University Press, Oxford, p 441
Pronin E (2006) Perception and misperception of bias in human judgment. Trends Cogn Sci 11:37
Stark A (2000) Conflicts of interest in American public life. Harvard University Press, Cambridge
Valsan R (2016) Fiduciary duties, conflict of interest and proper exercise of judgment. McGill Law J 62:1

Conflict of Laws

Amit M. Sachdeva
Ernst & Young LLP, Houston, TX, USA
New York University School of Law (NYU), New York, USA
London School of Economics (LSE), London, UK
The Hague Academy of International Law, The Hague, The Netherlands

Definition

The *Conflict of Laws*, also known as *Private International Law*, is that branch of the domestic

law of a State which deals with cases having a foreign element. (The word "State" has been used in the sense of a political system or a political subdivision having a separate and distinct legal system.)

All elements of any dispute can broadly be classified into those related to the parties and those related to the subject matter. When operating in a purely domestic scenario, i.e., when both/all the parties to the dispute hail from the same State and where the subject matter arises entirely within the territorial limits of the State, legal rights and obligations of the parties are required to be determined on the basis of substantive and procedural standards provided under the law of that State.

The position is, however, fundamentally different where, at least, one of the parties to the dispute is a foreign party or where the subject matter of the dispute arises, at least, partly outside the territory of the State whose courts are called upon to adjudicate it. In such situations, before proceeding to resolve the dispute and determine the rights and obligations of the parties, the court must first determine two important questions: (a) whether it has jurisdiction to resolve the dispute ("international jurisdiction question") and, (b) if so, by reference to the law of which State must the substantive rights and obligations be determined ("applicable law question").

These two questions, together with the question whether a judgment rendered by a foreign court may be recognized and/or enforced by the domestic court to which it is presented ("recognition question"), constitute, in very broad and general terms, the subject matter of the *Conflict of Laws.*

Scope, Purpose, and History

Narrower and Broader Scope

According to some scholars, the narrower understanding of the term "Conflict of Laws" dictates its use as limited only to the applicable law question inasmuch as while answering the applicable law question, the courts come across and resolve the "conflict" between "laws" of different States which may potentially govern the issue in dispute.

Parties to a transnational dispute, in most cases, stand to gain from whether a court exercises or declines to exercise jurisdiction and whether the substantive law of one State is preferred to the other. Similarly, a party with a favorable judgment handed down by a foreign court seeks to enforce its rights by means of recognition and enforcement of the (favorable) foreign judgment in other States than assuming the risk of litigating in each State. Thus, while "laws" of different States do not per se conflict – each being a comprehensive and complete system and regime in itself – the possibility of succeeding in a dispute induces significant disagreements between the parties about each of these three questions. The branch of domestic law of a State that concerns itself with answering any or all of these questions is, thus, called the *Conflict of Laws.* It is in this broader sense that the term is used here.

Need, Object, and Purpose

Since the earliest times, changes in kingship and integration of smaller kingdoms into empires resulted in (peaceful) coexistence of the newer with the older legal systems. Further, commercial dealings by men resulted in inevitable interaction between various legal systems which oftentimes gave rise to mutually incompatible rights and obligations.

In more recent times, with the growth in globalization and international trade, movement of persons, capital, labor, and resources has increased between States, resulting in courts being required to deal with matters involving foreign elements more often than not. This has necessitated the growth of *Conflict of Laws.*

The purpose of the conflicts rules is threefold:

(a) To define and confine the extent to which the laws of one State may extend extraterritorially.
(b) To resolve the disputes with a foreign element by exercise of international jurisdiction and by application of that law which results in substantial justice between the parties.
(c) To ensure that the litigation outcome is least distorted regardless of the State where the litigation is brought and that the substantive rights and obligations of the parties are adjudicated in the same manner.

In short, therefore, the object of the court in applying the conflicts rules is to reach at the outcome of a dispute which would have been reached in the jurisdiction and legal system to which it naturally and properly belonged.

History and Development

The history of *Conflict of Laws* goes back to a few centuries. The earliest history goes to the thirteenth-century universities in Italy during the times of the glossators and commentators, in particular Bartolus and Baldus.

A largely unbroken history of this branch dates back to the contributions of Charles Dumoulin's party autonomy and Bertrand d'Argentré's territorialism theories in France in the sixteenth century and Ulrich Huber in the Netherlands in the seventeenth century. The writings of the Dutch Huber were in the background of the "territorial sovereignty" thesis of the French Jean Bodin, whose thesis was strongly endorsed by the Dutch Hugo Grotius. Much of Huber's efforts and work focused on reconciling the application of foreign law with this overarching principle of territorial sovereignty, much in the interest of the Dutch being one of the leading international traders of that time.

The Dutch, unlike the Italians and the French scholars in the past, viewed the necessity of conflict arising out of the potential application of more than one competition legal system and are attributed the name "conflictus legum," directly translated as "Conflict of Laws." Huber, like Paul Voet, explained the resolution of this conflict sometimes resulting in the application of foreign law on the principle of "comity," a principle that apparently did not, in any manner, subdue the principle of territorial sovereignty. Huber laid the following three propositions which had unparalleled influence on the development of this branch of law in the common law:

(a) The laws of each State have effect within the territory of the sovereign, but not beyond.
(b) The laws of a State are binding on every person who is within, or enters, whether permanently or intermittently, the territory of the sovereign.
(c) On the ground of comity, a sovereign will allow rights acquired under the laws promulgated by other sovereigns to retain their force, so long as they do not prejudice the rights or powers of the sovereign or of its subjects.

These three propositions and, in particular, the third expatiated the rationale behind the application of foreign law by States as also carved out two important exceptions, which later more formally came to be referred as public policy and mandatory rules of the forum.

Subsequently, Friedrich Carl von Savigny, a German scholar, proposed to scientifically deal with the problem of Conflict of Laws and develop a scheme of yielding rules of universal application. He focused on the legal relationships and sought to seat each legal relationship to a jurisdiction to which it related. This resulted in neutral, ready-to-apply, and common system of conflicts rules that would potentially yield identical results regardless of the forum, forming the present-day system of *Conflict of Laws.*

Relationship with Public International Law

Public International Law constitutes the body of rules which govern the relationship between different States and normally sets the standard by which the validity of the conduct of a State is evaluated. This body of rules is a legal system in itself, independent of the municipal legal systems of the States. It may be noted however that there is a difference between the monist and the dualist perception of single or separate legal systems. Public International Law is embodied either in treaties or as customary international law.

On the other hand, *Conflict of Laws* is a part of the domestic law of the State that merely facilitates the domestic courts to resolve a particular category of disputes, i.e., those involving a foreign element. There is thus nothing more "international" about *Conflict of Laws* or *Private International Law* than that it governs issues of international jurisdiction, applicable law, and enforceability of foreign judgments in situations

that are not purely domestic. The conflicts rules are embodied in domestic statutes, court decisions, or jurisprudence and sometimes in treaties.

The Method of *Conflict of Laws*

The approaches and answers to the three questions forming the scope of the *Conflict of Laws* vary significantly. The following part offers a slightly detailed insight into the questions and offers an overview of some of the solutions that different States have to these questions.

International Jurisdiction Question

This question assumes particular significance because in an international context the party proposing to sue will, invariably, have at least two (and sometimes many) jurisdictions to choose from.

Since the law of the forum where the dispute is litigated determines its procedures as well as the applicability of mandatory rules and public policy, this question can at least in some cases be outcome determinative. The forum State's treaty network or reciprocity relations also have bearing on the enforceability of its judgments. Its importance can be gauged from the fact that in practice, the process of choosing the forum for litigation involves substantial time for a transnational lawyer.

While rules of international jurisdiction vary between States, there are some that are more general and common. Often, domestic rules on the point confer jurisdiction on their courts on the basis of the "link" that the defendant – the party being sued – has with the State. Some States accept "cause of action" as a basis for jurisdiction. In addition, most courts, in the present-day world, assume jurisdiction where conferred by the parties to the dispute.

Thus, where a defendant is domiciled or has the principal place of business in a State or the defendant has "minimum contacts" with the State, courts of that State have jurisdiction. This type of jurisdiction is called as "personal jurisdiction." Some States also assume jurisdiction on the basis of the nationality of the plaintiff, service of court process/writ, and the situation of defendant's property.

The court must also have "subject matter" jurisdiction that constitutes in some States an independent basis of jurisdiction and in some complements personal jurisdiction. Courts in several States entertain matters where the "cause of action" either wholly or partly arose within their territory.

Lastly, many courts, accepting the political and economic realities of this globalizing world, respect "consent-based" jurisdiction, i.e., where the parties either have a specific agreement that a chosen court shall deal with the dispute – called a "choice of court clause" – or, absent such agreement, the defendant does not, within a certain time, dispute or object to the jurisdiction of the court where the plaintiff has brought an action.

Common law countries significantly differ from the civil law countries in terms of "when" to exercise or decline jurisdiction. Common law courts have traditionally enjoyed much wider discretion in declining jurisdiction, even when all the prescribed requirements are otherwise satisfied. This discretion to decline the exercise of jurisdiction is exercised under different principles such as *forum non conveniens*, *lis alibi pendens*, foreign choice of court clauses, arbitration clauses, or abuse of court process (forum shopping). Apart from this, some common law courts are also known to issue "anti-suit injunctions" directed at the defendant (as opposed to the foreign State) preventing him from initiating or pursuing litigation in a foreign court, where the court is convinced that by resorting to another court, the defendant is likely to cause serious prejudice and manifest injustice to the plaintiff.

Applicable Law Question

The process of choosing the applicable law generally involves steps that are intricate and complex.

A typical choice of law analysis begins with the identification and characterization of the issue leading to the dispute. This requires the courts to identify, as a matter of domestic law, as narrowly and precisely as possible, one or more issues involved in the dispute. Once this is done, each

issue is "characterized" into one of the several known (or proposed) categories to which the choice of law rule must be applied. Thereafter, by applying the "connecting factor" prescribed by the concerned conflicts rule following from the characterization, the applicable law is determined.

Connecting factors or *points de rattachement* are those facts pertaining either to the parties or to the subject matter of dispute that tend to create nexus between the parties or dispute on one hand and a legal system on the other. In other words, connecting factors are those aspects which localize a controversy to one or more legal systems. As a generally accepted rule, the choice and prescription of connecting factors is done as a matter of the law of the forum, *lex fori.* Like any set of facts in all legal disputes, each fact carries different normative weight and depending upon the nature of the issue, different connecting factors may have lesser or greater "connecting" strength. For example, the domicile or residence of a contracting party is likely to be far less relevant while determining the governing law of contract than while determining the law governing the succession of movable property of a deceased.

While by no stretch of imagination exhaustive, the following is a list of conflicts rules relevant to different connecting factors that are most oftentimes used for determining the applicable law:

- *Lex fori*: law of the State where the forum/court is situated
- *Lex domicilii*: law of the State where the party is domiciled
- *Lex patriae*: law of the State of which a party is a national/citizen
- *Lex loci contractus*: law of the State where the contract was made
- *Lex loci solutionis*: law of the State where the contract is required to be performed
- *Lex delicti*: law of the State where a tort was committed
- *Lex loci celebrationis*: law of the State where a marriage was celebrated
- *Lex situs*: law of the State where a property is situated

An illustration would bring out the discussion more succinctly: Assume that XYZ A.G., a German holding company of a worldwide group, enters into a contract in Munich, Germany, with ABC Inc., a US corporation for sale and supply of x units of widgets, to its UK subsidiary, XYZ SubCo plc, for ultimate sale to consumers by the subsidiary in the UK. Now suppose that the goods are of an inferior quality and the action is brought in Germany. The rules by reference to which the German court in such a situation would determine the applicability of the law of one State in preference to the other is the subject matter of the applicable law question. The German court would first identify the issue between the parties: a contractual claim for sale of non-conforming goods. The court would then apply the conflicts rule appropriate to the concerned connecting factor. For example, if the German conflicts rule for determining the governing law in cases of contractual claims is that a contract is governed by the law of the place where it is made, the UK court would apply the German law.

In arriving at the applicable law, one comes across many difficult situations. Some of them are briefly highlighted here.

Characterization

Characterization of an issue into one as opposed to another may significantly change conflicts rule (and, consequently, the law) to be applied.

For instance, assume that under a legal system, two conflicts rules are known: (a) a contract is governed by the law of the State with which it is most closely connected, and (b) a tort claim is governed by the law where the tort is committed. Now suppose that Mr. X, a German national and domicile, purchased, while in Paris, a laptop from a French company, ABC, SA, for which Mr. X paid ABC in euros. Assume further that Mr. X carried the laptop with him to a seminar in India where the laptop short-circuited resulting in destruction of his research thesis. Mr. X now brings a claim against ABC in a court in Paris in France. If the short circuit is characterized as a contractual claim, then the French law (being the law to which the contract ex facie appears to be most closely connected) would apply. On the

other hand, if the French court were to characterize the claim as one arising in tort, it must apply the Indian law, being the law of the place where the tort was committed.

A related question which arises is: by reference to which law must the characterization of the issue be determined? Scholarship is fairly equally divided between those who advocate that characterization must be governed as a matter of domestic law of the State in which the court/forum is located (*lex fori)* and those who support the view that characterization itself must be governed by *lex causae.* Some recent thinkers urge for an internationalist approach to characterization.

Applicable Law: Internal or Conflicts Rules

The function of the *Conflict of Laws* however does not stop at identifying the "applicable law." Before the rights and obligations of the parties are determined by reference to the substantive and procedural rules prescribed in the municipal system of a State, the court must also determine the further question whether the applicable law identified by employing the appropriate conflicts rule refers directly to the internal law of that State or does it refer to the internal law as well as the conflicts rules of that State.

Counterintuitively, most legal scholarship and court decisions endorse the view that the applicable law includes not only the domestic law but also the conflicts rules of that State. This is founded on the reasoning that the dispute between the parties must be determined as if it was presented to the court whose law is determined to be the applicable law and then by resolving the dispute in the same manner as that court applying its own law (including the conflicts rules) would resolve.

Renvoi

Since the applicable law includes also the conflicts rules under the foreign law, it may potentially lead to a situation where on application by the court of the conflicts rules under the foreign law, it is determined that the dispute must be resolved by reference to the "law" of a third State or, in some cases, the law of the forum. "Law" of the third State or of the forum itself would constitute of domestic law as well as conflicts rules. The tendency of the court to apply the conflicts rule at this (and each subsequent) stage may result in either a ping-pong ball game or a relay race.

These "ping-pong ball" and "relay race" games are popularly known as *renvoi* and can potentially be avoided in many ways. One of the ways is by directly applying the internal law of the State whose law is determined to be the applicable law, i.e., applicable law may be understood as "applicable internal law." Another possibility could be to stop the reference to "law" as inclusive of conflicts rules after the first level, i.e., reference to "law" under the conflicts rules of the "applicable law" may be understood as the internal law, exclusive of the conflicts rules. Lastly, the court may consider the treatment of *renvoi* under the applicable law determined by applying the conflicts rules of the forum and act accordingly. Renvoi presents one of the most complicated and difficult issues in the area of *Conflict of Laws* and is laced with lack of predictability and judicial discretion.

Incidental Questions and Time Factor

Complications in determining the applicable law compound many folds when there is more than one interconnected issue. Referred to as the "incidental question" or "question préalable," it deals with ascertaining the legal system whose conflicts rules must determine the law governing the incidental question(s), i.e., whether the incidental question(s) must be determined by applying the conflicts rule of the law governing the main question or must the incidental question be determined by independently applying the conflicts rules of the forum.

Similarly, where either the conflicts rule or the content of the connecting factors or the applicable internal law undergoes change, serious difficulties arise about applying the altered rules to facts and circumstances prevailing before the change. Rare though, these questions have no easy answers.

Exclusion of Foreign Law

Courts however do not always apply the laws of other States, even when their conflicts rules lead to their application.

Called by different captions, a court would resist any temptation to apply foreign law, when doing so would either offend its public policy or conflict with its mandatory rules. Public policy or *ordre public* and mandatory rules represent those rules or values which, in a legal system, enjoy very high normative strength and, thus, have an overriding effect over any rule or norm to the contrary, including therefore those set by or under foreign law. This exception is broad based and controversial and one on which consensus is often difficult. It is pregnant with uncertainty, subjectivity, and judicial discretion.

Another exception that is widely accepted is enforcement of foreign penal and revenue laws. The rationale behind this exception is that penal and revenue laws give rise to "public," rather than "private," rights and liabilities and involve the assertion of sovereignty by a (foreign) sovereign over its subjects.

Recognition and Enforcement of Foreign Judgments

Under the traditional understanding of territorial sovereignty, a judgment rendered by a court has force within the territory of the State, but not beyond.

The "recognition" question constitutes the third and last scope of *Conflict of Laws*. Jurisprudentially, "recognition" and "enforcement" are different concepts. Enforcement requires positive action on behalf of a State to implement a foreign judgment, whereas recognition contemplates mere respect. A sword is to enforcement, what a shield is to recognition.

The recognition rules provide for the degree of finality a decision rendered by a court abroad enjoys in the State. Whether a decision rendered by a foreign court is conclusive of the rights and obligations of the parties and whether the parties are eligible, under certain circumstances, to re-litigate the issue before the courts of the concerned State are concerns that are addressed very differently by different States. The rules of different States are too disparate to admit any generalization. Suffice to say, absent an international treaty, the domestic rules that occupy the field reflect the peculiar historical development and geopolitical relations that the State whose court is petitioned to recognize a foreign judgment has with the State whose court has rendered that judgment.

The Hague Convention on the Recognition and Enforcement of Foreign Judgments in Civil and Commercial Matters, 1971, made some unsuccessful efforts to harmonize the law on the point and has entered into force for only five States. Efforts revived in the last decade of the previous millennium but failed to achieve an overambitious harmonization and settled with a far more humble treaty on choice of courts and recognition and enforcement of judgments rendered by the chosen court(s). In terms of this convention, a judgment rendered by a competent foreign court and which is subject to no further review is required to be recognized and enforced, unless the case falls under one of the exceptions stated in Article 5 of that convention.

Harmonization Efforts

As highlighted above, since the difference in the contents of laws of different States lies at the bottom of the Conflict of Laws, States have undertaken harmonization efforts in order to mitigate the difference(s) in the outcome.

General

These efforts can be divided in two categories: those seeking to harmonize the contents of the substantive internal law of the countries and those endeavoring synchronization of the conflicts rules across States. The latter is, relatively speaking, simpler since it involves synergizing only one branch among many braches of the domestic laws of different States. Several multilateral treaties have been entered into that deal with very specific aspects, such as carriage of goods by rail, carriage of passengers and luggage by road, marriage, divorce, adoption, minors' rights, nuclear energy, etc. Their formation and enforceability are, of course, matters of *Public International Law.*

European Union

Because of the peculiar supranational structure that the European Union is and because the EU

has the power to pass "regulations" in the area of *Conflict of Laws* that assume direct force as to both the text and effect in the Member States (with two exceptions), much of the conflicts rules have been harmonized by regulations at the EU level that have obliterated, as between the Member States, the operation of State conflicts rules.

In the EU, questions of jurisdiction pertaining to "civil and commercial matters" are governed by EC Regulation 44 of 2001, commonly called as "Brussels I Regulation." The EC Regulation contains well-defined and clear rules of jurisdiction, targeted to result in uniformity and predictability. The former is buttressed by the mandatory reference procedure to the European Court of Justice, which issues binding rulings interpreting the Regulation. Similarly, Brussels II bis Regulation, EC Regulation 2201 of 2003, provides for rules of jurisdiction in matrimonial matters and matters of parental responsibility.

The EU has also been largely successful in evolving a uniform system for determining the applicable law. The law on the point is basically contained in two EC Regulations, popularly called Rome I Regulation (EC Regulation 593 of 2008) and Rome II Regulation (EC Regulation 864 of 2007) providing for rules for determining the law applicable to contractual and non-contractual disputes, respectively.

A judgment rendered by one Member State of the EU is recognized and enforced by other Member States under the rules embodied in the two Brussels regulations and several other EC Regulations concerning recognition of judgments in specific cases, such as those pertaining to uncontested claims (EC Regulation 805 of 2004), insolvency (EC Regulation 1346 of 2000), etc.

The harmonization (being) brought among the Member States of the EU is, without doubt, the best model of harmonization yet known.

Conclusion

The conflicts inhering from the differences in legal systems have historically impeded businesses. The *Conflict of Laws* reflects one of the many where law, rather than facilitating economic activity, obstructs it. The impact (and sometimes deterrence) that the differences in legal systems on the economic efficiency is glaring and defies the most basic principles of *economics* and dictates further concerted action for harmonization.

References

Bariatti S (2011) Cases and materials on EU private international law. Hart, Oxford

Boele-Woelki K (2010) Unifying and harmonizing substantive law and the role of conflict of laws. Martinus Nijhoff, Boston

Bogdan M (2006) Concise introduction to EU private international law. Europa Law Publishing, Groningen

Briggs A (2008) The conflict of laws, 2nd edn. Oxford University Press, Oxford

Collins L (ed) (2012) Dicey, Morris and Collins on the conflict of laws, 15th edn. Sweet & Maxwell/Thomson Reuters, London

Esplugues C et al (2011) Application of foreign law. Sellier, Munich

Fawcett J (ed) (1995) Declining jurisdiction in private international law. Oxford University Press, Oxford

Hay P et al (2009) Comparative conflict of laws: conventions, regulations and codes. Thomson Reuters/Foundation Press, New York

Stone P (2010) EU private international law, 2nd edn. Edward Elgar, Northampton

Symeonides SC, Perdue WC (2012) Conflict of laws: American, comparative, international: cases and materials. West Thomson/Reuters, St. Paul

Consensus

Pavel Kuchař
Department of Economics and Finance, University of Guanajuato, DCEA-Sede Marfil, Guanajuato, GTO, Mexico
Facultad de Economía-División de Estudios de Posgrado, Universidad Nacional Autónoma de México, Ciudad de México, Mexico

Abstract

How does the process of consensus formation affect the accuracy and reliability of our knowledge? Cognitive and epistemic division of labor creates a problem of trust in the use

and application of knowledge. Consequently, the reliability of scientific consensus depends on whether the incentives, which the self-interested members of scientific communities face, are aligned in the right way.

Definition

Consensus is a conventional source of justified beliefs. In modern democratic societies, rational consensus, formed by means of free and open discussion, is a criterion of political legitimacy.

Introduction

Modern democracy as a form of government is a unique phenomenon, existing for only a short period of time in the history of our civilization. In this contribution, the role of consensus in directing the course of modern democracies is addressed. Public choice scholars have extensively studied the problem of amalgamating individual beliefs into aggregate social estimates for the purposes of legitimizing political authority. The aggregation of individual beliefs, however, is a specific example of a general question about the creation and use of knowledge. How does the process of consensus formation affect the accuracy and reliability of our knowledge? Without understanding how the division of labor brings together scientific communities and without understanding how these communities produce expert consensus, the criteria according to which we assess the accuracy and reliability of our warranted beliefs remain unclear.

According to literature, the reliability and accuracy of expert consensus depend on the nature of scientific institutions. In general, institutions are understood to be the rules of the game that determine the structure of payoffs for the agents involved. Given the cognitive and epistemic division of labor, expert consensus introduces a problem of trust. This problem is documented by empirical evidence that shows gaps between expert and nonexpert beliefs. If expert consensus results from the coordination of self-interested individuals, its reliability is dependent on whether the incentives, which the self-interested members of scientific communities face, are aligned in the right way.

Determining the right way to align incentives is ultimately an empirical problem. Therefore, the study of the economics of scientific knowledge through a comparative institutional analysis should be of interest to anyone who is curious about the role institutions play in the creation and use of knowledge in modern society.

Is There a Consensus Among Economic Experts?

If we want to know how the process of consensus formation influences the accuracy of our beliefs, we should perhaps first look into the state of expert consensus today. Some claim that the scope of consensus among economists is overwhelming; others, however, find this contention questionable. Generally, there seems to be a consensus regarding mainstream economic theory. With regard to economic policy, however, the attitudes among economic experts often differ.

According to Mark Blaug, "nothing like an overwhelming consensus has emerged from ... postwar economic methodology. But despite some blurring around the edges, it is possible to discern something like a mainstream view" (1992, p. 110). Today, this mainstream expert agreement can be summarized as follows: "Economists share the view that individuals are utility-maximizers, human wants are unlimited, and that mathematical modeling should be an important part of economic modeling" (May et al. 2013, p. 25). These are the basic building blocks of mainstream economic theory.

The scope of agreement among economists has been examined from the 1970s (Brittan 1973), and the surveys and expert polls seem to have established consensus on a range of economic questions. This consensus can be generally summarized with the following statement: "Price system or market is taken to be an effective and desirable social choice mechanism" (Frey et al. 1984, p. 994). In fact, Gordon and Dahl

(2013) found "no detectable systematic differences in views across departments, or across school of PhD." Moreover, they found "no evidence to support a conservative versus liberal divide" (p. 635). This would mean that regardless of where economists complete their education and regardless of their priors, they tend to agree on the core points of their discipline.

Expert consensus is, and has often been, challenged, however. Attitudes toward economic policy differ among male and female economists (May et al. 2013), among Democrat and Republican economists (Klein et al. 2013), and, in 1984, it was shown that expert opinion differed even among economists based in different countries: "The American, German, and Swiss economists tend to support more strongly the market and competition than their Austrian and French colleagues, who rather tend to view government interventions into the economy more favorably" (Frey et al. 1984, p. 994). These findings suggest that gender, political affiliation, and cultural and historical circumstances influence what kind of questions economists ask and what kind of answers they tend to give and agree on.

But why should we care if economists agree on anything? Presumably, the answer lies in the fact that economists are trained to take things which are not seen, at least not immediately, into account. The qualified point of view is then needed to uncover common fallacies. "What economists think, and whether there is consensus among economists, would not be a matter of concern if *beliefs* do not have a very strong effect on economic policy decisions and on the state of the economy" (Frey et al. 1984, p. 986 emphasis original). Expert opinion is a consequence of a particular kind of division of labor that takes place in modern democratic societies. As such, expert opinion is a source of warranted beliefs. The division of labor, however, introduces a problem of trust.

Can the Consensus Be Trusted?

A division of cognitive and epistemic labor has fostered the creation of knowledge by means of specialization and expertise. This division of labor, however, introduces problems of trust among the producers and consumers of expert knowledge. If the consumers of expert opinion do not trust its reliability, the expert consensus becomes inconsequential. Evidence shows that there is indeed a gap between expert and non-expert opinion.

In *The Republic of Science*, Michael Polanyi (1962, p. 471) argued,

> Scientific opinion is an opinion not held by any single human mind, but one which, split into thousands of fragments, is held by a multitude of individuals, each of whom endorses the others' opinion at second hand, by relying on the consensual chains which link him to all the others through a sequence of overlapping neighbourhoods.

There is a vast division of cognitive labor among scientists. Given his or her specialization, each expert looks at the world from a particular perspective. Each specialist, in turn, observes different aspects of reality. The deeper the specialization, the more diverse the observed aspects become. The essential difficulty consists in making sure these diverse observations are reliable and, above all, reconcilable with observations established by experts in distant neighborhoods of science.

Because of the division of cognitive labor, "nobody knows more than a tiny fragment of science well enough to judge its validity and value at first hand"; scientists have to "rely on views accepted at second hand on the authority of a community of people accredited as scientists" (Polanyi 1974, p. 173). Experts must mutually rely on the accuracy of peer review in scientific communities distant from their own. "If the aim of scientists is to maximize their knowledge of the world," writes Jesús Zamora Bonilla (2008, p. 4), "they need trust in the word of their colleagues, making science a collective enterprise."

On the one hand, the cognitive division of labor makes each scientist specialize in his or her area of expertise and build on the knowledge gained from exchanges and interactions with other experts, past and present. On the other hand, there is also epistemic division of labor.

Unlike the cognitive division of labor which results in specialization among scientists, epistemic division of labor separates the scientist from a nonscientist. At some point, the scientist had to make a decision to enter into the business of curiosity in the first place; it is a consequence of the epistemic division of labor that some people get to contribute to the production of expert consensus, while others become its mere consumers.

A pragmatist conception of democracy takes "the epistemic division of labor as one of the central features of effective and informed public deliberation"; there is a caveat, however: "Any such division of labor which produces epistemic gains will also produce deep asymmetries in the social distribution of knowledge" (Bohman 1999, p. 591). As a result, the body of scientific knowledge "can be received only when one person places an exceptional degree of confidence in another, the apprentice in the master, the student in the teacher, and popular audiences in distinguished speakers or famous writers" (Polanyi 1974, p. 220). Unless such a degree of confidence takes place, expert agreement loses its effect on public opinion.

Expert consensus has a strong effect on public opinion as long as it is relevant and credible. There are several reasons why expert consensus might be inconsequential, however. First, aspiring for abstract rigor, economic consensus might simply not be relevant for any of the problems nonexperts perceive as pressing (Mayer 1993; Šťastný 2010). Second, as Bryan Caplan (2007, p. 53) points out, even if generally relevant, the consensus may be perceived as biased. Caplan identified two main challenges to the objectivity of expert consensus:

> The first is *self-serving bias*. A large literature claims that human beings gravitate toward selfishly convenient beliefs. Since economists have high incomes and secure jobs, perhaps they are biased to believe that whatever benefits them, benefits all . . . The second doubt about economists' objectivity is less sordid but equally damaging: *ideological bias*. . . . A consensus of fundamentalists hardly inspires confidence. It sounds like an intellectual chain letter: Maybe each batch of graduate students was brainwashed by the previous generation of ideologues.

In his analysis, Caplan found a gap between what economists and laymen believe, suggesting that the exceptional degree of confidence in expert consensus is indeed missing. When identifying the sources of the gap, however, it became clear that the "self-serving and ideological bias *combined* cannot account for more than 20% of the lay/expert belief gap. The remaining 80% should be attributed to the experts' greater knowledge" (2007, p. 56 emphasis in original). The persistence of the gap must therefore be a consequence of some other cause. As Caplan hypothesizes, we "*turn off* our rational faculties on subjects where we don't care about the truth" (2007, p. 2 emphasis in original). Consequently, given the persistent ignorance, the expert consensus – even if generally unbiased and reliable – fails to have a strong impact on public opinion.

Another explanation of the gap suggests that although ignorance may be convenient, "the problem, it seems, is not that members of the public are unexposed or indifferent to what scientists say, but rather that they disagree about what scientists are telling them" (Kahan et al. 2011, p. 148). Even if the expert consensus was perceived as generally reliable, it could still fail to have a strong effect on public opinion as a result of a cultural cognition effect according to which "individuals systematically overestimate the degree of scientific support for positions they are culturally predisposed to accept" (pp. 166–167).

Indeed, as Gordon Gauchat (2012) shows, a "growing distrust in science in the United States has been driven by a group-specific decline among conservatives" (p. 179). Such a decline may reflect particular cultural, political, and ideological characteristics of the research agenda that the conservative group of expert-opinion consumers is not willing to take in. This explanation seems plausible, given the mostly progressive composition of the US expert group (Klein et al. (2013) shows that the proportion of Democrat economists to Republican economists is approximately three to one).

Note, for example, that unlike in the United States, there is a "a general tendency to the right among the Swedish social scientists" (Berggren et al. 2009, p. 2). If the political composition of

academia influences the approach to research, the situation Gauchat observed in the United States may actually run in reverse in countries like Sweden, where academia tends to be rather conservative as compared to the general public. In these cases, the progressive public may, in fact, display decreased perceived credibility of the scientific consensus arrived at by a more conservative cohort.

In short, there are two conditions needed for the scientific consensus to make a difference in the process of public deliberation. It must be reliable, and it must be credible. As long as it is reliable, the expert consensus has the potential to improve our knowledge of the world. As long as it is credible, the expert consensus has the power to influence public opinion. These two conditions are implied by the process of consensus formation which is inherently social and which therefore crucially depends on the nature of the institutional framework that supports it.

The Nature of Scientific Institutions

Institutions are the rules of the game; they determine the structure of payoffs for the agents involved in the game. If consensus formation is a social process and if expert consensus results from the coordination of self-interested members of the expert communities, the reliability of scientific consensus will depend on whether the incentives which the self-interested members of scientific communities face are aligned in the right way. It is then the analysis of the nature of scientific institutions that sets the stage for our understanding of the creation and use of knowledge in modern democratic societies.

Keeping in mind the cognition effect, let us assume that the influence of expert consensus is a function of its reliability: the more reliable and accurate the scientific consensus gets, the more credible it becomes. There are two essential notions to consider when determining the reliability of scientific consensus. First, scientists follow their self-interest; in this, they are no different from any other subject of economic analysis. Second, expert consensus, as a heuristic source of knowledge, induces conformity. Taking these assumptions into account implies that the reliability of expert opinion depends on particular properties of the scientific network that produced it. Let us look into these points further.

"Could it turn out," asks Philip Kitcher (1990, p. 6), "that high-minded inquirers, following principles of individual rationality, should do a poor job of promoting the epistemic projects of the community that they constitute?" The epistemic division of labor presumably sorts out people who are better at scientific research from people who are better equipped, for example to, say, start and run a business. This does not imply, however, that there is a hardwired feature in each and every scientist forcing him or her to pursue the discovery of truth at all costs.

It might not be in the best interest of a scientist to advance the stock of reliable knowledge. Brock and Durlauf (1999) emphasize this point: "Whereas the predominant themes in the philosophy of science . . . presumed . . . identical desire to find 'truth' . . . recent trends . . . have been concerned with the social context in which research is conducted" (p. 114). The literature has discerned that the social context of research produces motivations, which may compete with the presumed scientific urge to find truth.

Payoff structures determine the best course of actions for every scientist. When social context determines the structure of payoffs, it is reputation that scientists value highly. As Paul David (1998) argues, "A scientist working in a collegiate reputational reward system will consider the nearer-term reputational consequences of current actions (including expressions of scientific opinion), as well as considering long-term payoffs possibilities in the form of lasting fame for having gotten it right." In such a case, conformity may push against refining the body of reliable knowledge.

The success – or even survival – of scientists is, to a considerable extent, determined by their ability to publish in peer-reviewed journals. If getting it right coincides with getting it published, then "the implication is that referees act in the interests of science as a whole," as Frey (2003, p. 208) pointed out. But can the epistemic division of labor rely on the inherent goodwill of referees

involved in the peer review? Frey (2003, pp. 208–209) explains:

> Personal interests must also be expected to play a role. Many referees will be tempted to judge papers according to whether their own contributions are sufficiently appreciated and their own publications quoted. They carry, for instance, no costs when they advise rejection of a paper they dislike (e.g., because it criticizes their own work), even if they expect that it would be beneficial for economics as a discipline.

Economists may not question the generally followed assumptions of their discipline. In a situation when the problem at hand resembles Newtonian celestial dynamics, conforming to a mainstream consensus saves methodological effort. If, however, the sort of problems economists look into turns out to be better explained by some variant of generalized Darwinism, then following others in following Newton might lead the economist off the cliff – and into the land of the inconsequential.

According to Cass Sunstein, "following others can itself be seen as a heuristic, one that usually works well, but that also misfires in some cases" (2002, p. 38). Given the vast range of modern scientific knowledge, the consumers of expert consensus cannot help but follow expert opinion and trust that beliefs coming from the fields of science they are not acquainted with are well substantiated and properly justified. Expert consensus is a mental shortcut, a heuristic of consolidated opinion which, as J. S. Mill anticipated, "is salutary in the case of true opinions, as it is dangerous and noxious when the opinions are erroneous" (1859).

When conformity constitutes a rational course of action, "society can end up making large mistakes. . . . Social influences . . . threaten, much of the time, to lead individuals and institutions in the wrong directions." In such cases, "dissent can be an important corrective" (Sunstein 2002, p. 40). Dissenters, however, are often ignored and even ostracized, accused of destroying peaceful agreement for their own selfish motives. Furthermore, the dissenting opinion is often embedded within a system of language and reasoning that, from the perspective of the prevailing expert consensus, appears to lack rigor and preciseness. New ideas are substitutes for current prevailing thought. If reputation matters, and if the leading researchers are heavily invested in the prevailing consensus, plain dismissal of any dissent may occur.

When dissent does not pay off, originality does not take place. Dissenters, the individuals questioning the prevailing consensus, provide a valuable service that has a public-good character. By sharing their personal knowledge and pointing out where widely shared expert agreement runs short in terms of reliability, they contribute to the refinement of the body of scientific knowledge, benefiting all. Originality and dissent, however, are not absolutely valuable in themselves. They matter on the margin. Some questions of marginal analysis of dissent seem to follow: How much dissent is too much? And when does consensus become noxious?

The answer to these questions lies in the process of emergent consensus formation. An indispensable condition of producing reliable knowledge is an institutional mechanism, which ensures that the incentives scientists face align their self-interest with the general purpose of the epistemic division of labor. Through discovery and experimentation, an incentive-compatible institutional structure produces an emergent ratio of agreement and dissent. At this point, the production of scientific knowledge becomes a subject of comparative institutional analysis or, more broadly, economic analysis of scientific knowledge. "To the extent that we can make realistic presuppositions about human cognitive capacities and about the social relations found in actual communities of inquirers, we can explain, appraise and *in principle* improve our collective epistemic performance" (Kitcher 2002, p. 441 emphasis in original).

Given the agreement among economists about the price system being a desirable mechanism of social choice, it should be no surprise that one of the institutional arrangements suggested to improve the accuracy of scientific consensus is the prediction market, in other words, betting on beliefs. Prediction markets "doubtless have their limitations but they may be useful as a *supplement* to the other relatively primitive mechanisms for predicting the future like opinion

surveys, politically appointed panels of experts, hiring consultants or holding committee meetings" (Wolfers and Zitzewitz 2004, p. 125 emphasis added).

Robin Hanson, an advocate of prediction markets as an incentive-compatible institution of knowledge creation, claims that "betting prices . . . are a robustly accurate public institution estimating policy-relevant topics" (Hanson 2013, p. 156). The system of prediction markets supports dissent by rewarding out-of-favor beliefs, which turn out to be true, more than true beliefs that everybody supports. Such a system also seems to transparently draw the line between desirable and undesirable dissent because it encourages the well-informed to step forward and speak up and the ill-informed or dishonest to stay silent.

The scientific wager of Julian L. Simon and Paul Ehrlich on the implications of price theory is, after all, a well-known affair. The practice of betting on beliefs, however, does not come without problems. Today, prediction markets that aggregate information on questions of science or policy are disparaged in the same way life insurance – another form of betting on beliefs – used to be constrained by repugnance. With the exception of British prediction markets (generating odds on matters such as the likelihood of secession from the Eurozone or on the prognosis of the 2015 UK unemployment rate), betting on beliefs is mostly illegal.

Prediction market is but one of the institutional arrangements aggregating individual estimates into a composite indicator. Such an arrangement may provide some benefits, but does it outperform the current organization of science in producing justified beliefs? "Whether the rules according to which scientists compete for recognition among each other, and the rules that govern their competition for resources, are well aligned and whether they support or inhibit each other in promoting the growth of knowledge is an empirical matter" (Vanberg 2010, p. 43). Eventually, a comparative analysis should provide insights into the effects of diverse mutations of scientific institutions on the process of consensus formation.

Concluding Remarks

The epistemic performance of modern democratic societies depends on the division of labor in the creation of knowledge. At the same time, however, there is a gap between expert and nonexpert opinion, suggesting that the cognitive and epistemic division of labor creates a problem of trust in the use and application of knowledge. There are, in fact, several reasons why expert consensus may fail to impact public opinion: expert consensus may be irrelevant, biased, or simply opposed. In general, expert opinion fails to make a difference in the process of public deliberation when it is unreliable.

The reliability of expert consensus depends on the nature of scientific institutions. In other words, the reliability of scientific consensus depends on whether the incentives, which the self-interested members of scientific communities face, are aligned in the right way. The analysis of scientific institutions sets the stage for our understanding of the creation and use of knowledge in modern democratic societies. It is the empirical assessment of how different institutional arrangements perform in the social process of consensus formation that future research will have to address.

References

Berggren N, Jordahl H, Stern C (2009) The political opinions of Swedish social scientists. Finn Econ Pap 22(2):75–88

Blaug M (1992) The methodology of economics, or, how economists explain. Cambridge University Press, Cambridge/New York

Bohman J (1999) Democracy as inquiry, inquiry as democratic: pragmatism, social science, and the cognitive division of labor. Am J Polit Sci 43(2):590–607. https://doi.org/10.2307/2991808

Brittan S (1973) Is there an economic consensus?: an attitude survey. Macmillan, London

Brock WA, Durlauf SN (1999) A formal model of theory choice in science. Econ Theory 14(1):113–130

Caplan B (2007) The Myth of the rational voter: why democracies choose bad policies. Princeton University Press, Princeton

David PA (1998) Communication norms and the collective cognitive performance of "invisible colleges." In: Creation and transfer of knowledge. Springer, Heidelberg,

pp 115–163. Retrieved from http://link.springer.com/chapter/10.1007/978-3-662-03738-6_7
Frey BS (2003) Publishing as prostitution? – choosing between one's own ideas and academic success. Public Choice 116(1–2):205–223. https://doi.org/10.1023/A:1024208701874
Frey BS, Pommerehne WW, Schneider F, Gilbert G (1984) Consensus and dissension among economists: an empirical inquiry. Am Econ Rev 74(5):986–994. https://doi.org/10.2307/557
Gauchat G (2012) Politicization of science in the public sphere a study of public trust in the United States, 1974 to 2010. Am Sociol Rev 77(2):167–187. https://doi.org/10.1177/0003122412438225
Gordon R, Dahl GB (2013) Views among economists: professional consensus or point-counterpoint? Am Econ Rev 103(3):629–635. https://doi.org/10.1257/aer.103.3.629
Hanson R (2013) Shall we vote on values, but bet on beliefs? J Polit Philos 21(2):151–178. https://doi.org/10.1111/jopp.12008
Kahan DM, Jenkins-Smith H, Braman D (2011) Cultural cognition of scientific consensus. J Risk Res 14(2):147–174. https://doi.org/10.1080/13669877.2010.511246
Kitcher P (1990) The division of cognitive labor. J Philos 87(1):5–22. https://doi.org/10.2307/2026796
Kitcher P (2002) Contrasting conceptions of social epistemology. In: Brad Wray K (ed) Knowledge and inquiry: readings in epistemology. Broadview Press, Peterborough
Klein DB, Davis WL, Hedengren D (2013) Economics professors' voting, policy views, favorite economists, and frequent lack of consensus. Econ J Watch 10(1):116–125
May A, McGarvey MG, Whaples R (2013) Are disagreements among male and female economists marginal at best?: a survey of AEA members and their views on economic policy. Contemp Econ Policy. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/coep.12004/full
Mayer T (1993) Truth versus precision in economics. E. Elgar, Aldershot/Brookfield
Mill JS (1859) On liberty. In: Essays on politics and society, vol XVIII. University of Toronto Press, Toronto
Polanyi M (1962) The republic of science: its political and economic theory. Minerva 38(1):1–21
Polanyi M (1974) Personal knowledge: towards a post-critical philosophy. University of Chicago Press, Chicago
Šťastný D (2010) The economics of economics: why economists aren't as important as garbagemen (but they might be). Instituto Bruno Leoni/CEVRO Institute and Wolters Kluwer, Turin/Prague
Sunstein CR (2002) Conformity and dissent. Law School, University of Chicago, Chicago. Retrieved from http://www.law.uchicago.edu/files/files/34.crs_.conformity.pdf
Vanberg VJ (2010) The "science-as-market" analogy: a constitutional economics perspective. Constit Polit Econ 21(1):28–49. https://doi.org/10.1007/s10602-008-9061-5
Wolfers J, Zitzewitz E (2004) Prediction markets. J Econ Perspect 18(2):107–126. https://doi.org/10.2307/3216893
Zamora Bonilla JP (2008) The elementary economics of scientific consensus. Theor Rev Teoría Hist Fundam Cienc 14(3):461–488

Consequentialism

David Moroz
Champagne School of Management, Groupe ESC Troyes, France

Abstract

Adopting a consequentialist approach requires knowing the whole set of consequences of an action. If we assume the future is radically uncertain, which may be argued for several different reasons, this approach can be used only in an explicative way and not in a predictive one.

Definition

Consequentialism can be defined as an instrumentalist approach to ethics that consists in evaluating a given system through its resulting effects (Pettit 2003; Blackburn 2008), i.e., through the maximization of gains and the minimization of losses it enables (Baggini and Fosl 2007).

Several forms of consequentialism can be identified (Baggini and Fosl 2007; Thiroux and Krasemann 2012), among which are the different kinds of utilitarianism. What enables the comparison between the different kinds of consequentialism can be said to be (i) the set of principles that are adopted to define positive as well as negative consequences and (ii) the choice of the agents, in a given system, that should benefit from positive consequences.

Consequentialism and Knowledge

From a methodological point of view, and if we want to embrace wider issues than the mere sphere of economic science, we cannot disconnect the

issues of consequentialism from the issues of knowledge. Indeed, outside of its ethical dimension, consequentialism is related to the ability to define causal relationships to enable the prediction of the results emerging from a given system (Thiroux and Krasemann 2012): choosing between alternative systems on the basis of expected results requires evaluating ex ante the relevance of said results. Such an issue is problematic for any consequentialist approach, even in the case where one individual determines for herself what positive and negative consequences are, i.e. even in the case where the subjectivity of individual preferences is respected.

To be able to predict with certainty the results of a given system, three conditions are necessary (Faber and Proops 1993): (i) the knowledge of its initial conditions; (ii) the absence of novelty; and (iii) the knowledge of its laws. Yet, three arguments enable us to explain that the future is radically uncertain: (i) the impossibility of predicting the evolution of knowledge; (ii) the cognitive limits of the analysis of complexity; and (iii) the impossibility of defining the limits of our ignorance.

Concerning the first argument, Popper (1957) explains that "if there is such a thing as growing human knowledge, then we cannot anticipate today what we shall know only tomorrow." So it is impossible to assert that the last two conditions necessary for the prediction of a system – absence of novelty and knowledge of its laws – can be met.

The second argument relies on the limits of the analysis of so-called complex phenomena. As explained by Hayek, our ability to analyze complex phenomena is necessarily limited because of the impossibility for the human brain, as for any "apparatus of classification" (Hayek 1952), to analyze a system more complex than itself. Consequently, the theories we build can only be *ceteris paribus* constructions (Lachmann 1978) that cannot take account of the totality of variables that may influence a phenomenon.

Concerning the third argument, it can be found partly in Kirzner (1992) with the concept of limit of ignorance. Let us suppose it is possible to differentiate among the set of future events, the predictable events and the unpredictable events. Drawing the frontier between these two sets of events would require knowing the limit of our ignorance, which is antithetical.

The concept of radical uncertainty does not mean that the future cannot be imagined, but that it cannot be known before its time (Lachmann 1978). In the opposite situation, the notion of novelty would be meaningless: novelty or surprise cannot exist in a world where the future is already known (see Shackle 1972, 1979).

This impossibility of building a causal relationship concerning the future does not, however, prevent building any causal relationship. It is here that the concept of time relativity defined by Hicks (1979), which enables us to differentiate a fixed past time and an evolving future time, takes on its full meaning. Past time is a closed set of events that can be analyzed following a deterministic approach (De Uriarte 1990), whereas future time is an open set of events, the result of which cannot be predicted.

Conclusion

As Hodgson (1996) explains, even if the world is determinist, we should consider it unpredictable. Indeed, if we adopt the deterministic approach, it is impossible to explain either the possibility of changing the future (Lawson 1988; Davidson 1996), of creating it (Shackle 1972; Lachmann 1977; Loasby 1991, 1999), or the existence of choice and action (Von Mises 1963; Shackle 1972; Hodgson 1996), i.e., free will (De Uriarte 1990; Hodgson 1996).

The rub is that if we assume the world as unpredictable, so as to consider the possibility for individuals to choose and to create, then resorting to any consequentialist approach to establish normative prescriptions becomes highly problematic.

Cross-References

► Choice Under Risk and Uncertainty

References

Baggini J, Fosl PS (2007) The ethics toolkit: a compendium of ethical concepts and methods. Blackwell, Oxford
Blackburn S (2008) The Oxford dictionary of philosophy, 2nd edn. Oxford University Press, Oxford
Davidson P (1996) Reality and economic theory. J Post Keynes Econ 18(4):479–508
De Uriarte B (1990) On the free will of rational agents in neoclassical economics. J Post Keynes Econ 12(4):605–617
Faber M, Proops JLR (1993) Evolution, time, production and the environment, 2nd edn. Springer Verlag, Berlin Heidelberg
Hayek FA (1952) The sensory order: an inquiry into the foundations of theoretical psychology. University of Chicago Press, Chicago
Hicks J (1979) Causality in economics. Basil Blackwell, Oxford
Hodgson GM (1996) Economics and evolution: bringing life back into economics. University of Michigan Press, Ann Arbor
Kirzner IM (1992) The meaning of market process: essays in the development of modern Austrian economics. Routledge, London
Lachmann LM (1977) Capital, expectations, and the market process: essays on the theory of the market economy. Subsidiary of Universal Press Syndicate, Kansas City
Lachmann LM (1978) An Austrian stocktaking: unsettled questions and tentative answers. In: Spadaro LM (ed) New directions in Austrian economics. Sheed Andrew and McMeel, Kansas City, pp 1–18
Lawson T (1988) Probability and uncertainty in economic analysis. J Post Keynes Econ 11(1):38–65
Loasby BJ (1991) Equilibrium and evolution: an exploration of connecting principles in economics. Manchester University Press, Manchester
Loasby BJ (1999) Knowledge, institutions and evolution in economics. Routledge, London
Pettit P (2003) Consequentialism. In: Darwall S - (ed) Consequentialism. Blackwell, Oxford, pp 95–107
Popper KR (1957) The poverty of historicism. Routledge & Kegan Paul, London
Shackle GLS (1972) Epistemics and economics: a critique of economic doctrines. Cambridge University Press, Cambridge
Shackle GLS (1979) Imagination and the nature of choice. Edinburgh University Press, Edinburgh
Thiroux JP, Krasemann KW (2012) Ethics: theory and practice, 11th edn. Pearson Prentice Hill, Upper Saddle River
Von Mises L (1963) Human action: a treatise on economics, 4th edn. Fox & Wilkes, San Francisco

Further Reading

Auyang SY (1998) Foundations of complex-system theories in economics, evolutionary biology, and statistical physics. Cambridge University Press, Cambridge
Hayek FA (1978) The pretence of knowledge. In: Hayek FA (ed) New studies in philosophy, politics, and economics, and the history of ideas. Routledge & Kegan Paul, London, pp 23–24
Hodgson GM (2004) Darwinism, causality and the social sciences. J Econ Methodol 11(2):175–194
Lawson T (1997) Economics and reality. Routledge, London
Salmon WC (1998) Causality and explanation. Oxford University Press, New York
Shackle GLS (1967) Décision, Déterminisme et Temps. Dunod, Paris

Constitutional Economics

▶ Constitutional Political Economy

Constitutional Evolution in Ancient Athens

George Tridimas
Department of Accounting, Finance and Economics, Ulster University Business School, Belfast, Northern Ireland

Definition

The present essay reviews recent political economy research in the policy-making institutions of the direct democracy of ancient Athens, 508–322 (all dates BCE), their origins, and evolution. The historical events, especially tensions between rich and poor and existential external threats, which led to the emergence of the Athenian democracy, are first described. The institutions of decision-making, assembly, magistrates, and courts, are then discussed along with various reforms in response to changing circumstances. The essay ends with some reflections on the nature of the direct democracy, its legitimatization, and the internal consistency of its institutions in comparison to its modern representative analog.

The Birth of Democracy

Constitutional economics raises time-invariant questions. Inquiring how historical societies

addressed constitutional building, architecture of governance, and rights of governed, not only allows to make sense of the past but also helps to understand present arrangements, as well as showing the strengths and limitations of theoretical models. This is even more so for democracy with its key elements of freedom and political equality. It is this realization which has inspired a vibrant branch of research by both political economists and historians into the political institutions of ancient Athens, their origins, evolution, and performance. The present essay surveys this work and reflects on the constitutional evolution of ancient Athens.

The constitutional development of ancient Athens is the story of the emergence of direct participatory democracy (albeit for adult men only) where political events and institutional building were inextricably linked (Aristotle (1984) for the original account; Raaflaub et al. (2007) for debate on the origins of the Athenian democracy; Cartledge (2016) for birth and tribulations of ancient democracy compared to modern). Table 1 summarizes the main developments.

In the Archaic Times (750–500), Athens was governed by the nine archons appointed from the members of the noble, landed, elite, who were responsible for religious, military, civic affairs, and recording laws. After serving one-year terms, the archons were appointed for life as members of the Council of *Areopagus* with the authority to oversee laws and magistrates and conduct trials. In 594, following a period of internal social conflicts, Solon, an aristocrat, was appointed as lawgiver and introduced a series of institutional and economic changes. Inscribed and publically displayed, the laws keynoted that governance secretly managed by the aristocracy had ended. Among other issues, Solon conditioned appointment to public office on wealth (valued in annual agricultural production) rather than hereditary birthright, with the highest income classes enjoying access to more powerful offices and the lowest (the landless) altogether excluded. He also granted all citizens the right to participate in the assembly (which was more like a consultative body as it was not empowered to make binding decisions) and act as prosecutors in criminal trials, and introduced accountability of magistrates. In 546, Peisistratus established tyranny, meaning extra-constitutional dispensation rather than oppressive government, which lasted until 510 (see Fleck and Hanssen (2013) on the transition from tyranny to democracy). In the ensuing contest for power between two aristocrats

Constitutional Evolution in Ancient Athens, Table 1 Timeline of the Athenian democracy

750–500	**Archaic Athens** Main government bodies: Nine archons selected from the aristocracy; ex-archons appointed to the Areopagus overseeing laws and magistrates and conducted trials
594	Solon, the lawgiver, introduced a wealth-based political dispensation
546–510	Tyranny of Peisistratus and his son Hippias
510	Hippias expelled
508–404	**Classical Athens. Fifth Century Democracy**
508–507	Democracy established: Cleisthenes reforms of citizenship and council of five hundred. Ostracism introduced
490–479	Persian wars. 480: Athenian victory at Marathon; 490: Athenian victory at Salamis
487	Selection of nine archons by lot
462	Powers of Areopagus removed. Introduction of pay for court service
451	Pericles' law restricts citizenship to those whose both parents were Athenians
431–404	*Peloponnesian War, Athens V Sparta*
415–413	Sicilian expedition of Athenian navy. Syracuse and Sparta defeat Athens
411	Democracy overthrown by oligarchic coup
410	Democracy restored by the Athenian navy
405	Defeat of Athens at Aegospotami
404	Athenian defeat and surrender; Tyranny of the Thirty
404–322	**Classical Athens. Fourth Century Democracy**
403	Democracy restored
403/402	Introduction of pay for attending the assembly
ca 402	The assembly voted that new laws would be made by boards of legislators appointed by lot
ca 355	*Theoric* (festival money) fund formalized
338	Athens & Thebes defeated in Chaeronea by Philip of Macedon
322	End of the Athenian Democracy after defeats by Macedon in the Lamian War

Isagoras and Cleisthenes, the former prevailed. Then, in an unprecedented move, Cleisthenes allied himself to the demos, the common people, proposing reforms that would extend political rights. Isagoras turned to oligarchic Sparta, the strongest military power at the time for help. Spartan forces occupied Athens and expelled Cleisthenes. However, when they tried to dissolve the governing council and impose Isagoras, the Athenian demos rose up forcing the Spartans to leave. Cleisthenes was recalled in 508 and introduced a series of fundamental constitutional reforms 508/7 that gave birth to the direct participatory democracy.

The Institutions of the Athenian Direct Democracy

In the first instance, the reforms reconstituted the rules for citizenship, the idea that all locally born free men within a city-state had equal political rights and enjoyed legal protections, regardless of wealth, birth, education, or any other factor (Hansen (1999) offers an extensive treatment of the operation of the Athenian democracy; Ober (2008) focuses on aggregation of private knowledge through democratic institutions; Tridimas (2011) and Lyttkens (2013) analyze the institutions of democratic governance in the light of political economy). Citizenship rights were extended to all adult resident males, a move equivalent to egalitarian enfranchisement (citizenship rights were, however, limited by Pericles in 451 to those whose both parents were Athenians). Cleisthenes then divided the citizens into three geographical sections, Urban, Inland, and Coast, and further divided each geographical section to ten parts and a total of 139 demes (local communities) whose membership was hereditary. A lottery allocated the resulting 30 groups to 10 new tribes ("*phylae*"), so that each new tribe included groups from each one of the three geographical sections with their different traditions and economic bases. Each tribe was a microcosm of the citizenry and shared the same interests with the rest of the tribes. The new structure translated into a more effective and successful Athenian military (see Pritchard 2015 on the contribution of democracy to the military success of Athens).

Next, the reforms established the assembly of (male) Athenian citizens as the principal decision-making body, responsible for all domestic and external policy issues, including military issues, foreign alliances, tax, expenditure, infrastructure works, and public festivals. All adult male Athenians had the right to vote and address the assembly. Out of a population of perhaps 60,000 male citizens in the fifth century, the quorum stood at 6000. The quorum remained the same during the fourth century when as a result of disease, war losses, and famine population fell to 30,000. In the fifth century, the assembly met 10 times a year, but gradually the number increased to 40 in the second half of the fourth century (Tridimas (2017) on the choice of frequency of voting). Contrary to modern practice, policy measures were proposed by individuals in their capacity as citizens rather than office holders. Neither political parties nor executive offices as known today existed. It is for this reason that the ancient authors never refer to anything resembling the office of president or prime minister. Although all Athenians had the right to propose policy measures and address the assembly, in practice a small number of political leaders, statesmen, dominated its debates. Decisions were taken by majority (see Pitsoulis (2011) for the origins of majority voting). Voting took place by show of hands; if the show was unclear, then the chairing officers counted hands.

Cleisthenes set up a new body the Council of Five Hundred (fifty from each tribe) selected annually by lot, with responsibilities to prepare the agenda of the assembly and carry out the day-to-day administration of Athens. Members of each tribe chaired the administration of Athens for one-tenth of the year. Every day at sunset, the 50 councilors of the presiding tribe selected by lot a chairman, who was the head of the Athenian state for a night and a day. The members of the Council met every working day and received a compensation for their services, so that poorer citizens could afford to take time off their daily work and serve in public office. Councilors too voted by show of hands. A man could serve in the Council only

twice in his lifetime, implying a considerable turnaround in the office and, as a consequence, a large number of Athenians with some experience of the affairs of the state.

In 501, the board of the Ten Generals was introduced to serve as commanders of the army and navy for a year. In a special assembly meeting, candidates were proposed from each tribe, but had to be voted in by a majority of the voters of all tribes, so a General was not the political representative of his tribe; from 440 at least one General was elected from all tribes. In contrast to other office-holders who were subject to term limits, generals could be reelected.

The *Heliaia* Court of 6000, or "People's Court," first set up by Solon to hear appeals against the decisions of the officials, had its powers enlarged in 462 and became the most important court with wide responsibilities. Every year 6000 citizens (600 from each tribe) not in debt to the state were selected by lot to serve as jurors. Each day jurors were allocated to cases by lot and sat in sessions with a jury size of 501 or bigger as the case may be (201 minimum). There was no public prosecutor all parties appearing before the court, citizens who brought a charge, the magistrates preparing and presiding over a case and the jurors who decided were ordinary citizens without any legal training. Contrary to assembly votes, jury decisions were taken by secret ballot. Pay for jurors was introduced most probably in 462 thought to be equal to half the average wage. From the fifth century, the court tried both civil and penal cases, checked the eligibility and conduct of public officers (but not competence), and conducted trials for treason and corruption (Karayiannis and Hatzis (2012) for social norms and the rule of law; Carugati et al. (2015) for the decentralized legal order of Athens).

In addition, another six hundred, or so, magistrates were appointed annually by lot to serve the polis. They typically worked in boards of ten members and carried specific administrative tasks, including inspection of markets, supervision of public works, judicial administration, collection of state revenues, etc.

Cleisthenes also introduced ostracism (banishment) whereby the demos in a secret ballot decided whether to banish a leading individual for a period of 10 years, but without any further criminal or financial penalties. It was conceived as a mechanism to stop destructive violent conflict for power by political leaders and to defend the democracy. It was used sparingly with ten attested ostracisms during the period 507–416 and none afterward, although the procedure was not abolished (see Tridimas (2016) for a game theoretic analysis).

Military success followed setting Athens in a glorious trajectory. In 490, the hoplites army of landowner–farmer citizens pushed back the invading Persians in the battle of Marathon. In 483/2, a rich silver vein was discovered in Southern Attica. The demos voted to use the windfall to build a navy instead of distributing it equally among the citizenry, which was a common practice at the time (Tridimas 2013). The fleet triumphed against the Persians in the 480 sea battle of Salamis. Athens then transformed to a sea power precipitating a shift in the internal balance of political power against the large and mid-size landowners who were the backbone of the land forces, and in favor of the poor class of the landless. They had found gainful employment as rowers in the newly built fleet, and on the basis of their new strength, they gained access to all political offices. By the mid-fifth century, direct democracy for the Athenian male citizens was fully functioning.

After the final defeat of the Persians in 479, Athens pursued an offensive war heading the Delian League of Greek islands and coastal city-states. In 478, the League was transformed into the Athenian hegemony, where the allies were paying tribute to Athens for protection by her navy (see Rhodes (2013) for the Athenian public finances). Athens reached the peak of its power, excelling in public monument building and as the intellectual and cultural center of the time.

War, Constitutional Reforms, Recovery, and End of Democracy

Intense rivalry between Athens and Sparta, the traditional Greek military power, led to the

Peloponnesian War, 431–404, spreading to the rest of Greece, Asia Minor, and the Greek colonies of Southern Italy and Sicily. Convulsions followed the 413 destruction of the Athenian fleet in Sicily. In 411, the democracy was overthrown by an oligarchic regime restricting full political rights to 5000 men only. Four months later, in 410, after defeating the Spartan fleet the navy reinstated the democracy. In 404, Athens was finally defeated. With Spartan backing, Athens was ruled by a cruel 30-member strong oligarchic commission, known as the "Thirty Tyrants." The democrats regrouped, and in 403 after some fierce fighting, they expelled the Tyrants and restored democracy.

A number of institutional changes followed. Blaming the "demagogues" for misleading the demos in the Peloponnesian War, the powers of the assembly were curtailed. Instead of the assembly, a special board of legislators chosen by lot from the same panel of 6000 jurors of the People's Court was appointed to pass laws describing "general norms without limit of duration." The assembly still voted decrees and decided foreign policy (Schwartzberg (2004) on the sovereignty of the demos and legal change). In what can be thought as an early type of judicial constitutional review, the courts acquired more powers through deciding a lawsuit alleging an unconstitutional proposal. Accordingly, they could nullify assembly measures deemed contrary to the laws, unfavorable to the interests of the people, or procedurally invalid, and punish their proposers (Lyttkens et al. 2017). Nevertheless, it differs from modern judicial review for the latter is carried out by professional judges, while the Athenian one was entrusted to ordinary folk, amateurs who had no formal legal training. Thus, the demos exercised its rule in both the assembly and the courts (Hansen 1999).

Funding of a variety of state functions was fixed by law diminishing the discretionary power of the Assembly to allocate public expenditures. The new elected offices of the treasurer of the military fund, the board of the *theoric* fund to manage festival money, and the controller of the finances were also established. Supervision of the administration of the laws was transferred to the Areopagus. Finally, pay for the first 6000 citizens attending the Assembly at the average daily wage was introduced. A further development of the fourth century was the relative decline of Generals as political leaders in comparison to the fifth century when Generals dominated assembly deliberations and provided policy direction. In the fourth century, without holding any formal office, a number of self-selected orators rose to prominence debating in the assembly and the courts, while the Generals focused on more narrow military matters.

Like all democracies, Athens redistributed away from the rich by taxing them and mandating them to finance various public services, known as liturgies (Kaiser 2007; Lyttkens 2013; Tridimas (2015b) on rent-seeking). Although not manifested through political parties, for there were none in the direct democracy, the division between the rich and the poor was clear and was evident in foreign policy. The rich, who carried the burden of war finance, and the farmers anxious when leaving their lands to fight, favored peace. But the poor with less property, the prospect of gainful employment in the fleet and land allotments abroad if victorious, favored war, while also worried about the abolition of democracy if Athens' enemies prevailed.

Over the fourth century, the Athenian economy recovered after the defeat, and eventually thrived based on international trade buttressed by non-discriminatory laws that focused on transactions rather the origin (Athenian or foreign) of the transactors and a stable regulatory framework (Ober (2015) and Bresson (2016) for the performance of the Athenian economy; Bitros and Karayiannis (2010) on moral norms of Athens, Bergh and Lyttkens (2014) on institutional quality; Economou and Kyriazis (2017) on the evolution of private property rights). Tax revenues rose substantially (Kyriazis 2009). New alliances were formed and several wars were fought against rival Greek city-states, but Athens failed to achieve the earlier supremacy. In the 338 battle of Chaeronea, Philip of Macedon inflicted a heavy defeat on an Athenian–Theban alliance forcing them to join the Macedonian alliance that led the invasion of Persia under Alexander the Great.

Following the death of Alexander the Great in 322, Athens fought to break away from the Macedon-led alliance. It suffered a double defeat in the sea battle of Amorgos and the land battle of Crannon. An oligarchy was then established with Macedonian support. The democracy ended as the oligarchy abolished the courts, disenfranchised the poor (limiting the franchise to 9000 property owners), stopped pay for public service and assembly participation, abolished large numbers of liturgies, and reduced assembly meetings to ceremonial functions (Tridimas 2015a).

Reflections on the Nature of the Athenian Democracy

Democracy means rule of the demos, the many. Nevertheless, its Athenian essence was not just numerical, the majority versus the minority of citizens, but also economic, that is, democracy connoted rule by the "middling" and the "poor," meaning all those who unlike the "rich" had to work for a living (Patriquin 2015). In explaining the emergence of the Athenian democracy, ancient historians focus on identifying "historically contingent factors," the weight of important events as in 594, 508, and 462. On the contrary, economists search for reasons why the enfranchised elite may accept democracy which works against their privileged access to power. Fleck and Hanssen (2006) show that democratic institutions resolve time inconsistency problems, and thus it increases investment. In the broken territory of Athens, more suitable for olive oil than wheat production, monitoring the input of captive labor was next to impossible. The only way for the elite to ensure that nonelite engage in productive investment, so that it generates the necessary surplus to finance defense, is when the nonelite enjoy long-run security of their properties. The latter is achieved when the nonelite have full political rights so that they can vote for policies conducive to their interests. McCannon (2012) explains that democracy was beneficial to both the elite and the nonelite of Athens. The Athenian elite experienced significant wealth volatility, which implied that the children of the current generation of the elite faced a considerable risk of suffering wealth losses and losing their privileged position. Democracy with its egalitarian principle of access to power not conditional on wealth and its redistributive impact provided an insurance mechanism against the risk of loss of wealth and political standing.

The historical narrative indicates that the establishment of the Athenian democracy was a gradual and cumulative process of institutional building over a long period in response to changing political, military, and economic circumstances. Starting with the reforms of Solon, which based eligibility for public office on wealth holdings rather than aristocratic birthright, and including an interval of tyranny; the process moved on with the reforms of Cleisthenes, which enfranchised all Athenian males, introduced ostracism as a peaceful way to resolve conflicts via electoral means, then opened up public office to the poorer citizens. Further, democracy–building encompassed appointment to public office by lot, payment of a service fee, so that citizens could afford time off their production activities for public service, and further changes in the fourth century which included payment for assembly attendance, conceptualization of the difference between permanent laws and assembly decrees, elevation the courts to an effective veto player by granting them the power to check the decisions of the assembly, and introduction of new treasury offices. The process bears remarkable similarities to Congleton's (2011) theory of the emergence of western democracy and representative government from the industrial revolution to the twentieth century, that is, slow and piecemeal, building on established arrangements and more peaceful than violent. More importantly, there was nothing teleological about the establishment of the democracy, it was not inevitable, and as described despite its comparative longevity 507–322, it did not survive.

Table 2 compares the institutions of the Athenian democracy with those of a stylized modern democracy.

The idea of citizenship was central to the polis. Citizenship conferred political and civic rights and economic benefits, and was strictly regulated.

Constitutional Evolution in Ancient Athens, Table 2 Comparison of Athenian and modern democracies

	Athenian democracy	Modern democracy[a]
Policy making	Direct and participatory: The assembly chose policy	Indirect: Citizens vote for party candidates, who then choose policy
Legitimacy[b] of government	Equality of opportunity that citizens may occupy public office	Citizens consent to be governed by those who win an electoral majority
Political parties	Absent	Fundamental players in electoral competition
Chief executive	Not important	President or Prime Minister
Frequency of decision-making	Frequent assembly meetings per year	Periodic elections
Voting rule	Simple majority	Varies from majoritarian to proportional representation
Appointment to public office	By lot Election limited to a few offices	Election
Justice and administration of the state	Ordinary citizens took turns to serve as jurors and carry out administrative tasks	Professional judiciary Professional bureaucracy
Policy review and accountability of officials	Popular court – juries of ordinary citizens	Courts – professional judges/legal experts

[a]Stylized description of modern representative government
[b]Legitimacy: the decision taker is recognized to have the right to do what he does

Citizens were expected to be active participants in the polis from fighting wars to deciding public policy and holding public office regularly. It marked a clear break with the eastern empires preceding the Greek city-states, where ordinary people were subjects of kings and obliged to pay tribute to them. On the other side of the spectrum, the premise of the ancient democracy was that all citizens have equal opportunities to serve in public office. This is unlike modern representative democracies which are predicated on the principle that citizens have equal opportunities to consent to how they are governed (Manin 1997).

The true hallmark of democracy was appointment to office by lot (also called sortition), rather than voting. The luck of the draw eliminated the advantages that the rich elite had in contesting elections for office (Taylor 2007; Tridimas 2012; Lyttkens 2013). The lot nullified the policy-making privileges of public offices formerly controlled by the aristocratic class, and by randomizing appointment to office, it also randomized the distribution of rents from holding office. Of course, sortition cannot select political leaders with the qualifications to govern, nor can exclude incompetent individuals from office. The Athenians were aware that not all citizens were suitable for positions of responsibility. Thus, only offices that did not require specialized expertise were filled by lot. To put it another way, the Athenians accepted that the acts of boards filled by sortitioned members were invariant to the composition of those boards. On the other hand, offices requiring leadership skills, like the position of the general who had to be trusted in leading to battle, were filled by elections. The sortitioned public postholders served annual terms implying that considerable rotation took place.

The Athenian institutions of direct democracy, decision-making in the assembly by majority voting, appointment to office by lot, accountability to the courts consisting of ordinary members of the demos instead of professional legal experts, absence of political parties, and lack of professional expertise in public administration, comprised an integral structure, consistent with each other and mutually reinforcing. No part of the constitutional package could operate independently of the rest. When voters decide directly on every single policy issue which can be introduced by any willing citizen (a) simple majority suffices, (b) there is no need for political

parties to bundle issues in party platforms, and (c) the search for voting formulas other than simple majority to aggregate votes is negated. Further, in the absence of elections for candidates to office, the courts with juries randomly drawn from the citizenry ensured accountability. Equally, when direct democracy empowers citizens to occupy public office through sortition, implying that any citizen might hold office, and annual rotation, implying that every citizen might hold office at some time, (a) political parties can no longer secure rents for their members, (b) no classes of professional politicians or bureaucrats emerge, and (c) citizens have an incentive to learn about the running of the state and "all things political."

Conclusion

It is clear from the above brief account that ancient and modern democracy share the basic principle of people's rule and equality in the sense of one man one vote, but in truth little else. Assembly debate to decide public policy and sortition to administer the state, rather than voting for candidates and party political intermediation, were the fundamentals of the former. Undoubtedly, the direct and the representative democracies share common values and principles, but conceptually and practically they are apart.

Cross-References

- ▶ Constitutional Political Economy
- ▶ Voting Power Indices

References

Aristotle (1984). The Athenian constitution. London: Penguin Classics. (translated by P. J. Rhodes).

Bergh A, Lyttkens CH (2014) Measuring institutional quality in ancient Athens. J Inst Econ 10:279–310

Bitros GC, Karayiannis AD (2010) Morality, institutions and the wealth of nations: some lessons from ancient Greece. Eur J Polit Econ *26*:68–81

Bresson A (2016) The making of the ancient Greek economy: institutions, markets and growth in the city–states. Princeton University Press, Princeton. (translated by S Rendall)

Cartledge P (2016) Democracy. A life. Oxford University Press, Oxford

Carugati F, Hadfield G, Weingast B (2015) Building legal order in ancient Athens. J Legal Anal 7:291–324

Congleton RD (2011) Perfecting parliament: Constitutional reform and the origins of western democracy. Cambridge University Press, New York

Economou EML, Kyriazis N (2017) The emergence and the evolution of property rights in ancient Greece. J Inst Econ 13:53–77

Fleck RK, Hanssen FA (2006) The origins of democracy: a model with applications to Ancient Greece. J Law Econ 49:115–146

Fleck RK, Hanssen FA (2013) How tyranny paved the way to democracy: the democratic transition in ancient Greece. J Law Econ 56:389–416

Hansen MH (1999) The Athenian democracy in the age of demosthenes. Structure, principles and ideology. Bristol Classical Press, London

Kaiser BA (2007) The Athenian trierarchy: mechanism design for the private provision of public goods. J Econ Hist 67:445–480

Karayiannis AD, Hatzis AN (2012) Morality, social norms and the rule of law as transaction cost-saving devices: the case of ancient Athens. Eur J Law Econ 33:621–643

Kyriazis NK (2009) Financing the Athenian state: public choice in the age of Demosthenes. European Journal of Law and Economics 27:109–27

Lyttkens CH (2013) Economic analysis of institutional change in Ancient Greece: politics, taxation and rational behaviour. Routledge, Abingdon

Lyttkens CH, Tridimas G, Lindgren' A (2017) Making direct democracy work. An economic perspective on the *graphe paranomon* in ancient Athens. Available at http://swopec.hhs.se/lunewp/abs/lunewp2017_010.htm

Manin B (1997) The principles of representative government. Cambridge University Press, Cambridge

McCannon BC (2012) The origin of democracy in Athens. Rev Law Econom 8:531–562

Ober J (2008) Democracy and knowledge. Princeton University Press, Princeton and Oxford

Ober J (2015) The rise and fall of classical Greece. Princeton University Press, Princeton and Oxford

Patriquin L (2015) Economic equality and direct democracy in ancient Athens. Palgrave MacMillan, New York

Pitsoulis A (2011) The egalitarian battlefield: Reflections on the origins of majority rule in archaic Greece. European Journal of Political Economy 27:87–103

Pritchard DM (2015) Democracy and war in ancient Athens and today. Greece and Rome 62:140–154

Raaflaub KA, Ober J, Wallace RW (eds) (2007) Origins of democracy in Ancient Greece. University of California Press, Berkely

Rhodes PJ (2013) The organization of Athenian public finance. Greece and Rome 60:203–231

Schwartzberg M (2004) Athenian democracy and legal change. Am Polit Sci Rev 98:311–325
Taylor C (2007) From the Whole Citizen Body? The Sociology of Election and Lot in the Athenian Democracy. Hesperia 76:323–346
Tridimas G (2011) A political economy perspective of direct democracy in ancient Athens. Constit Polit Econ 22:58–72
Tridimas G (2012) Constitutional choice in ancient Athens: the rationality of selection to office by lot. Constit Polit Econ 23:1–21
Tridimas G (2013) Homo oeconomicus in ancient Athens. Silver bonanza and the choice to build a navy. Homo Oeconomicus 30:14–162
Tridimas G (2015a) War, disenfranchisement and the fall of the ancient Athenian democracy. Eur J Polit Econ 31:102–117
Tridimas G (2015b) Rent seeking in the democracy of ancient Greece. In: Hillman AL, Congleton RD (eds) The Elgar companion to the political economy of rent seeking. Edward Elgar Publishing, Cheltenham, pp 444–469
Tridimas G (2016) Conflict, democracy and voter choice: a public choice analysis of the Athenian ostracism. Public Choice 169:137–159
Tridimas G (2017) Constitutional choice in Ancient Athens: the evolution of the frequency of decision making. Constit Polit Econ 28:209–230

Constitutional Political Economy

Stefan Voigt
Institute of Law and Economics, University of Hamburg, Hamburg, Germany

Abstract

Economists used to be interested in analyzing decisions assuming the rules to be given. Scholars of Constitutional Political Economy (CPE) or constitutional economics have broadened the scope of economic research by analyzing both the choice of basic rule systems (constitutions) as well as their effects using the standard method of economics, i.e., rational choice.

Synonyms

Constitutional economics

Definition

Constitutional Political Economy analyzes the choice of constitutions as well as their effects by drawing on rational choice.

Introduction

Buchanan and Tullock (1962, p. vii) define a constitution as "... a set of rules that is agreed upon in advance and within which subsequent action will be conducted." Although quite a few rule systems could be analyzed as constitutions under this definition, the most frequently analyzed rule system remains the constitution of the nation state. Two broad avenues in the economic analysis of constitutions can be distinguished: (1) the normative branch of the research program, which is interested in legitimizing the state and the actions of its representatives, and (2) the positive branch of the research program, which is interested in explaining (a) the outcomes that are the consequence of (alternative) constitutional rules and (b) the emergence and modification of constitutional rules.

Normative Constitutional Political Economy

Methodological Foundations

Normative CPE deals with a variety of questions, such as: (1) How should societies proceed in order to bring about constitutional rules that fulfill certain criteria, like being "just"? (2) What contents should the constitutional rules have? (3) Which issues should be dealt with in the constitution – and which should be left to post-constitutional choice? (4) What characteristics should constitutional rules have?, and many more. James Buchanan, one of the founders of CPE and the best-known representative of its normative branch, answers none of these questions directly but offers a conceptual frame that would make them answerable. The frame is based on social contract theory as developed most prominently by Hobbes. According to Buchanan (1987,

p. 249), the purpose of this approach is justificatory in the sense that "it offers a basis for normative evaluation. Could the observed rules that constrain the activity of ordinary politics have emerged from agreement in constitutional contract? To the extent that this question can be affirmatively answered, we have established a legitimating linkage between the individual and the state."

The value judgment that a priori nobody's goals and values should be more important than anybody else's is the basis of Buchanan's entire model. One implication of this norm is that societal goals cannot exist. According to this view, every individual has the right to pursue her own ends within the frame of collectively agreed upon rules. Accordingly, there can be no collective evaluation criterion that compares the societal "is" with some "ought" since there is no such thing as a societal "ought." But it is possible to derive a procedural norm from the value judgment just described. Buchanan borrowed this idea from Knut Wicksell (1896): Agreements to exchange private goods are judged as advantageous if the involved parties agree voluntarily. The agreement is supposed to be advantageous because the involved parties expect to be better off with the agreement than without it. Buchanan follows Wicksell who had demanded the same evaluation criterion for decisions that affect more than two parties, at the extreme an entire society. Rules that have consequences for all members of society can only be looked upon as advantageous if every member of society has voluntarily agreed to them. This is the Pareto criterion applied to collectivities. Deviations from the unanimity principle could occur during a decision process on the production of collective goods, but this would only be within the realm of the Buchanan model as long as the constitution itself provides for a decision rule below unanimity. Deviations from the unanimity rule would have to be based on a provision that was established unanimously.

Giving Efficiency Another Meaning

Normative CPE thus reinterprets the Pareto criterion in a twofold way: It is not outcomes but rules that, in turn, lead to outcomes which are evaluated using the criterion. The evaluation is not carried out by an omniscient scientist or politician but by the concerned individuals themselves. To find out what people want, Buchanan proposes to carry out a consensus test. The specification of this test is crucial as to which rules can be considered legitimate. In 1959, Buchanan had factual unanimity in mind and those citizens who expect to be adversely affected by some rule changes would have to be factually compensated. Later, Buchanan seems to have changed his position: Hypothetical consent deduced by an economist will do to legitimize some rule (see, e.g., Buchanan 1977, 1978, 1986). This position can be criticized because a large variety of rules seem to be legitimizable depending on the assumptions of the academic who does the evaluation. Scholars arguing in favor of an extensive welfare state will most likely assume risk-averse individuals, while scholars who argue for cuts in the welfare budgets will assume that people are risk neutral.

Buchanan and Tullock (1962, p. 78) introduced the veil of uncertainty, under which the individual cannot make any long-term predictions as to her future socioeconomic position. As a result, unanimous agreement becomes more likely. John Rawls' (1971) veil of ignorance is more radical because the consenting individuals are asked to decide *as if* they did not have any knowledge on who they are. Both veils assume a rather curious asymmetry concerning certain kinds of knowledge: On the one hand, the citizens are supposed to know very little about their own socioeconomic position, but on the other, they are supposed to have detailed knowledge concerning the working properties of alternative constitutional rules (Voigt (2013) summarizes veilonomics).

Every society must decide on which actions are to be carried out on the individual level and which ones are to be carried out on the collective one. To conceptualize this decision problem, Buchanan and Tullock introduce three cost categories: *External costs* are those costs that the individual expects to bear as a result of the actions of others over which she has no direct control. *Decision-making costs* are those which the individual expects to incur as a result of her own participation

in an organized activity. The sum of these two cost categories is called *interdependence costs*. Only if their minimum is lower than the costs under private action will society opt in favor of collective action.

After having decided that a certain activity is to be carried out collectively, societies need to choose decision-making rules that are to be used to make concrete policy choices. The more inclusive the decision-making rule, the lower the external costs an individual can expect to be exposed to. Under unanimity, they are zero. Decision-making costs tend, however, to increase dramatically the larger the number of individuals required to take collective action. Buchanan and Tullock claim that "for a given activity the fully rational individual, at the time of constitutional choice, will try to choose that decision-making rule which will *minimize* the present value of the expected costs that he must suffer" (1962, p. 70).

Positive Constitutional Economics

Positive CPE can be divided into two parts: On the one hand, it is interested in explaining the outcomes that result from alternative rule sets. On the other, it is interested in explaining the emergence and modification of constitutional rules. Following Buchanan and Tullock (1962), the genuine contribution of CPE to economics should consist in endogenizing constitutional rules. To date, analysis of the effects of constitutional rules has, however, played a more prominent role. This is why we begin by summarizing the research on the effects of constitutions.

The Effects of Constitutional Rules

Quite a few empirical studies confirm that constitutional rules can cause important differences, only some of which can be mentioned here (Voigt (2011) is a more complete survey). Conceptually, it makes sense to distinguish between the catalogue of basic rights on the one hand and the organization of the various state representatives *(Staatsorganisationsrecht*) on the other. Here, we confine ourselves to the latter and begin by looking at the effects of electoral rules and the form of government as these two aspects of constitutions have been analyzed the most. We continue with federalism and direct democracy. Other organizational aspects that deserve to be analyzed include uni- versus bicameral legislatures and the independence that so-called non-majoritarian institutions such as central banks enjoy. Further, analysis is confined to democracies, as analyzing the effects of electoral rules makes little sense with regard to autocracies.

Electoral Rules

According to Elkins et al. (2009), only some 20% of all constitutions contain explicit provisions regarding the electoral system to be used for the choice of parliamentarians. For two reasons, we nevertheless begin this brief survey with electoral systems: (1) The definition of constitutions introduced above does not imply that only constitutions "with a capital C" can be recognized and (2) the effects of these rules seem to be more significant and robust than those of other constitutional traits.

A distinction is sometimes made between electoral rules and electoral systems. *Electoral rules* refer to the way votes translate into parliamentary seats. The two most prominent rules are majority rule (MR) and proportional representation (PR). *Electoral systems* include additional dimensions such as district size and ballot structure. Although theoretically distinct, these dimensions are highly correlated empirically: Countries relying on MR often have a minimum district size (single-member districts) and allow voting for individual candidates. Countries that favor PR often have large districts and restrict the possibility of deviating from party lists.

Duverger's (1954) observations that MR is conducive to two-party systems whereas under PR more parties are apt to arise have been called "Duverger's law" and "Duverger's hypothesis," respectively, documenting the general validity ascribed to them. Research into the economic consequences of electoral systems began much later. It has been argued (Austen-Smith 2000) that since coalition governments are more likely under PR than under MR, a common pool problem among governing parties will emerge. Parties participating in the coalition will want to please

different constituencies, which explains why both government spending and tax rates are higher under PR than under MR.

In their survey of the economic effects of electoral *systems*, Persson and Tabellini (2003) also investigate district size and ballot structure. Suppose single-member districts are combined with MR, which is often the case empirically. In this situation, a party needs only 25% of the national vote to win the elections (50% of half of the districts; Buchanan and Tullock 1962). Contrast this with a single national district that is combined with PR. Here, a party needs 50% of the national vote to win. Persson and Tabellini (2000) argue that this gives parties under a PR system a strong incentive to offer general public goods, whereas parties under MR have an incentive to focus on the swing states and promise policies that are specifically targeted at the constituents' preferences.

The effects of differences in ballot structure are the last aspect of electoral systems to be considered. Often, MR systems rely on individual candidates, whereas proportional systems rely on party lists. Party lists can be interpreted as a common pool, which means that individual candidates can be expected to invest less in their campaigns under PR than under MR. Persson and Tabellini (2000) argue that corruption and political rents should be higher, the lower the ratio between individually elected legislators and legislators delegated by their parties.

Persson and Tabellini (2003) put these conjectures to an empirical test on the basis of up to 85 countries and a period of almost four decades (1960–1998). They find the following effects: (1) In MR systems, central government expenditure is some 3% of GDP lower than in PR systems. (2) Expenditures for social services are some 2–3% lower in MR systems. (3) The budget deficit in MR systems is some 1–2% below that of systems with PR. (4) A higher proportion of individually elected candidates is associated with lower levels of perceived corruption. (5) Countries with smaller electoral districts tend to have less corruption. (6) A larger proportion of individually elected candidates is correlated with higher output per worker. (7) Countries with smaller electoral districts tend to have lower output per worker.

Blume et al. (2009a) replicate and extend Persson and Tabellini's analysis, finding that with regard to various dependent variables, district magnitude and the proportion of individually elected candidates is more significant – both substantially and statistically – than the electoral rule itself.

Form of Government: Presidential Versus Parliamentary Systems

In parliamentary systems, the head of government is subject to a vote of no confidence by parliament and can, hence, be easily replaced by the legislature. In presidential systems, the head of government is elected for a fixed term and cannot be easily replaced. Conventional wisdom holds that the degree of separation of powers is greater in presidential than in parliamentary systems as the president does not depend on the confidence of the legislature. Persson et al. (1997, 2000) argue that it is easier for legislatures to collude with the executive in parliamentary systems, which is why they expect more corruption and higher taxes in those systems than in presidential systems. They further argue that the majority (of both voters and legislators) in parliamentary systems can pass spending programs whose benefits are clearly targeted at themselves, implying that they are able to advance their interests to the detriment of the minority. This is why they predict both taxes and government expenditures to be higher in parliamentary than in presidential systems.

Empirically, Persson and Tabellini (2003) find that (1) government spending is some 6% of GDP lower in presidential systems. (2) The size of the welfare state is some 2–3% lower in presidential systems. (3) Presidential systems seem to have lower levels of corruption. (4) There are no significant differences in the level of government efficiency between the two forms of government. (5) Presidential systems appear to be a hindrance to increased productivity, but this result is significant at the 10% level only.

These results are impressive and intriguing. Although presidential systems exhibit lower government spending and suffer less from corruption than parliamentary systems, the latter enjoy

greater productivity. In a replication study, Blume et al. (2009a) find that the results are not robust, even to minor modifications. Increasing the number of observations from 80 to 92 makes the presidential dummy insignificant in explaining variation in central government expenditure. This is also the case as soon as a slightly different delineation of presidentialism is used. If the dependent variable is changed to total (instead of central) government expenditure, the dummy also becomes insignificant.

The differences in these results clarify an important point: To date, many of the effects supposedly induced by constitutional rules are not very robust; they crucially hinge upon the exact specification of the variables, the sample chosen, the control variables included, and so on. This suggests that further research needs to be as specific as possible in trying to identify possible transmission channels and to take into consideration the possibility that small differences in institutional details can have far-reaching effects.

Vertical Separation of Powers: Federalism

The vertical separation of powers has been analyzed in economics as "fiscal federalism." Scholars of this approach largely remain within the traditional model, i.e., they assume government to be efficiency maximizing. They then ask on what governmental level public goods will be (optimally) provided, taking externalities explicitly into account. This approach need not concern us here because it does not model politicians as maximizing their own utility (for surveys, see Inman and Rubinfeld 1997; Oates 1999, 2005).

The economic benefits of federalism are thought to arise from the competition between constituent governments (i.e., from noncooperation); its costs are based in the necessity of cooperating on some issues (i.e., from cooperation). Hayek (1939) argued long ago that competition between governments reveals information on efficient ways to provide public goods. Assuming that governments have incentives to make use of that information, government efficiency should be higher in federations, ceteris paribus. In Tiebout's (1956) famous model, the lower government levels compete for taxpaying citizens, thus giving lower governments an incentive to cater to these citizens' preferences.

Turning to possible costs of federal constitutions, Tanzi (2000) suspects that government units that provide public goods will be insufficiently specialized because there might be too many of them. Also, federal states need to deal with a moral hazard problem that does not exist in unitary states. The federal government will regularly issue "no-bail-out clauses" but they will not always be credible.

What can we learn from existing empirical studies? For a long time, the evidence concerning the effects of federalism on overall government spending was mixed. Over the last decade, though, this has changed. Rodden (2003) shows that countries in which local and state governments have the competence to set the tax base, total government expenditure is lower. Feld et al. (2003) find that more intense tax competition leads to lower public revenue.

Based on principal component analysis, Voigt and Blume (2012) conclude that institutional detail clearly matters. They find that with regard to a number of dependent variables (budget deficit, government expenditures, budget composition, government effectiveness, and two measures of corruption), frequently used federalism dummies turn out to be insignificant in explaining variation, whereas particular aspects of federalism are significant, some of them very strongly so. One problem for this research strategy is the small number of observations as there are only some 20 federal states countries in the world. One possible way of circumventing this problem is to draw on case studies instead of econometric estimates.

Direct-Democratic Versus Representative Institutions

Representatives of normative CPE ask what rules the members of a society could agree on behind a veil of uncertainty. If they agree on a democratic constitution, they would further have to specify whether and to what extent they want to combine representative with direct-democratic elements. To make an informed decision, the citizens would be interested to know

whether direct democracy institutions display systematic effects.

In most real-world societies, representative and direct democracy are complementary because it is impossible to vote on all issues directly. Among direct democracy institutions, referendums are usually distinguished from initiatives. The constitution can prescribe the use of referendums for passing certain types of legislation, in which case agenda-setting power remains with the parliament but citizen consent is required. Initiatives, in contrast, allow citizens to become agenda setters: The citizens propose legislation that will then be decided upon, given that they manage to secure a certain quorum of votes in favor of the initiative. Initiatives can aim at different levels of legislation (constitutional versus ordinary legislation), and their scope can vary immensely (e.g., some constitutions prohibit initiatives on budget-relevant issues).

These institutions can mitigate the principal-agent problem between citizen voters and politicians. If politicians dislike being corrected by their citizens, direct democracy creates incentives for politicians to implement policies that are closer to the preferences of the median voter. Initiatives further enable citizens to unpack package deals resulting from logrolling between various representatives. If citizens like only half of a deal made by politicians, they can start an initiative trying to bring down the other half.

To date, most empirical studies on the effects of direct democracy institutions have focused on within country studies, in particular with regard to the USA and Switzerland. Most studies have confirmed theoretical priors. Matsusaka (1995, 2004), e.g., finds that US states with the right to an initiative have lower expenditures and lower revenues than states without that institution. Feld and Savioz (1997) find that per capita GDP in Swiss cantons with extended democracy rights is some 5% higher than in cantons without such rights. Frey and coauthors argue that one should investigate not only the outcomes that direct-democratic institutions produce but also the political processes they induce (e.g., Frey and Stutzer 2006). Indeed there is some evidence that citizens in countries with direct democracy institutions have better knowledge about their political institutions (Benz and Stutzer 2004) and are more interested in politics in general (Blume and Voigt 2014).

Blume et al. (2009b) is the first cross-country study to analyze the economic effects of direct democracy. They find a significant influence of direct-democratic institutions on fiscal policy variables and government efficiency, but no significant correlation between direct-democratic institutions and productivity or happiness. Institutional detail matters a great deal: While mandatory referendums appear to constrain government spending, initiatives seem to increase it. The actual use of direct-democratic institutions often has more significant effects than their potential use, implying that – contrary to what economists would expect – the direct effect of direct-democratic institutions is more relevant than its indirect effect. It is also noteworthy that the effects are usually stronger in countries with weaker democracies.

Constitutional Rules as Explanandum

Procedures for Generating Constitutional Rules

Constitutional rules can be analyzed as the outcome of the procedures that are used to bring them about. Jon Elster's (1991, 1993) research program concerning CPE puts a strong emphasis on hypotheses of this kind. He inquires about the consequences of time limits for constitutional conventions, about how constitutional conventions that simultaneously serve as legislature allocate their time between the two functions, about which effects the regular information of the public concerning the progress of the constitutional negotiations has, and about how certain supermajorities and election rules can determine the outcome of conventions.

Ginsburg et al. (2009) is a survey of both the theoretical conjectures and the available empirical evidence. They expect constitution-making processes centered on the legislature to be associated with greater post-constitutional legislative powers whereas constitution-making processes centered on the executive to be negatively correlated with such powers. Interestingly, the first half of the conjecture is not supported by the data whereas the second half is. In a book-length work on the

life span of constitutions, Elkins et al. (2009) report that public involvement in constitution-making is indeed correlated with a longer constitutional life spans.

The Relevance of Preferences and Restrictions for Generating Constitutional Rules

Procedures are a modus of aggregating inputs and can therefore never produce constitutional rules by themselves. It is thus only a logical step to analyze whether a set of potentially relevant variables can explain the choice of certain constitutional rules. There are good reasons – confirmed by some empirical evidence – to assume that (1) the individual preferences of the members of constitutional conventions will directly enter into the deliberations and that (2) the preferences of all the citizens concerned will be recognized in the final document in quite diverse ways. This would mean that rent-seeking does play a role even on the constitutional level, a conjecture that is often rejected by representatives of normative CPE.

McGuire and Ohsfeldt (1986, 1989a, b) have tried to explain the voting behavior of the Philadelphia delegates and of the delegates to the 13 states ratifying conventions that led to the US constitution. Their statistical results show that, ceteris paribus, merchants, western landowners, financiers, and large public-securities holders supported the new constitution, whereas debtors and slave owners opposed it (1989a, p. 175).

Explicit Versus Implicit Constitutional Change

The two approaches toward constitutions as explananda just sketched are rather static. A third approach focuses on explaining modifications of constitutions over time. Constitutional change that results in a modified document will be called explicit constitutional change here whereas constitutional change that does not result in a modified document – i.e., change that is due to a different interpretation of formally unaltered rules – will be called implicit constitutional change.

One approach toward explaining long-run explicit constitutional change focuses on changes of the relative bargaining power of organized groups. Due to a comparative advantage in using violence (see North 1981), an autocrat is able to establish government and secure a rent from that activity. As soon as an (organized) group is convinced that its own cooperation with the autocrat is crucial for the maintenance of the rent, it will seek negotiations with the autocrat. Since the current constitution is the basis for the autocrat's ability to appropriate a rent, the opposition will strive to change it. In this approach, bargaining power is defined as the capability to inflict costs on your opponent. The prediction of this approach is that a change in (relative) bargaining power will lead to modified constitutional rules (Voigt 1999).

Future Directions

CPE has been developing very fast over the last 20 years. The availability of very large and detailed datasets has definitely contributed to this development. But many questions still need additional research. Here are some of the relevant issues.

Regarding the effects of alternative constitutional rules, we know very little regarding both human rights and constitutional amendment rules. Both can, however, be extremely important. Regarding the effects of alternative constitutional rules in general, establishing causality is a challenge. There is little variation of constitutional rules over time which makes it difficult to establish causality. A possible solution is to change the level of analysis and to leave the nation-state level and move down to the communal level where both the number of observations and the number of changes in the rules are higher.

By now, very detailed datasets concerning the de jure content of constitutions are available. But it is well known that the de facto situation in many countries does not exactly reflect the de jure contents of a constitution. One important area for future research in CPE would, thus, try to identify the determinants for the divergence between de jure provisions and de facto reality.

There are only a handful of papers trying to identify the determinants of the choice of concrete constitutional provisions such as the choice between a presidential and a parliamentary form

of government (Hayo and Voigt 2013; Robinson and Torvik 2008). It is almost certain that there will be more results on other constitutional rules soon. In a broader context, identifying the determinants that lead to changes from an autocratic to a democratic constitution – and vice versa – remains highly desirable.

References

Austen-Smith D (2000) Redistributing income under proportional representation. J Polit Econ 108(6): 1235–1269

Benz M, Stutzer A (2004) Are voters better informed when they have a larger say in politics? – evidence for the European Union and Switzerland. Public Choice 119:31–59

Blume L, Voigt S (2014) Direct democracy and political participation. Forthcoming

Blume L, Müller J, Voigt S, Wolf C (2009a) The economic effects of constitutions: replicating – and extending – Persson and Tabellini. Public Choice 139:197–225

Blume L, Müller J, Voigt S (2009b) The economic effects of direct democracy – a first global assessment. Public Choice 140:431–461

Buchanan JM (1977) Freedom in constitutional contract – perspectives of a political economist. Texas A&M University Press, College Station/London

Buchanan J (1978) A contractarian perspective on anarchy. In: Pennock JR, Chapman JW (eds). Anarchism. Anarchism, New York, pp 29–42

Buchanan JM (1987) The constitution of economic policy. Am Econ Rev 77:243–250

Buchanan J (1986) Political economy and social philosophy In: ders.; Liberty, market and state – political economy in the 1980s. New York: Wheatsheaf Books, pp 261–74

Buchanan JM, Tullock G (1962) The calculus of consent – logical foundations of constitutional democracy. University of Michigan Press, Ann Arbor

Duverger M (1954) Political parties: their organization and activity in the modern state. Wiley, New York

Elkins Z, Ginsburg T, Melton J (2009) The endurance of national constitutions. Cambridge University Press, Cambridge

Elster J (1991) Arguing and bargaining in two constituent assemblies, The Storrs Lectures

Elster J (1993) Constitution-making in Eastern Europe: rebuilding the boat in the Open Sea. Public Adm 71(1/2):169–217

Feld LP, Savioz M (1997) Direct democracy matters for economic performance: an empirical investigation. Kyklos 50(4):507–538

Feld L, Kirchgässner G, Schaltegger C (2003) Decentralized taxation and the size of government: evidence from Swiss state and local governments, CESifo working paper 1087, Dec

Frey B, Stutzer A (2006) Direct democracy: designing a living constitution. In: Congleton R (ed) Democratic constitutional design and public policy – analysis and evidence. MIT Press, Cambridge, pp 39–80

Ginsburg T, Elkins Z, Blount J (2009) Does the process of constitution-making matter? Ann Rev Law Sci 5:5.1–5.23

Hayek F (1939) Economic conditions of inter-state federalism. New Commonw Quart S2:131–149

Hayo B, Voigt S (2013) Endogenous constitutions: politics and politicians matter, economic outcomes don't. J Econ Behav Organ 88:47–61

Inman R, Rubinfeld D (1997) Rethinking federalism. J Econ Perspect 11(4):43–64

Matsusaka J (1995) Fiscal effects of the voter initiative: evidence from the last 30 years. J Polit Econ 102(2):587–623

Matsusaka J (2004) For the many or the few. The initiative, public policy, and American Democracy. The University of Chicago Press, Chicago

McGuire RA, Ohsfeldt RL (1986) An economic model of voting behavior over specific issues at the constitutional convention of 1787. J Econ Hist 46(1):79–111

McGuire RA, Ohsfeldt RL (1989a) Self-Interest, agency theory, and political voting behavior: the ratification of the United States Constitution. Am Econ Rev 79(1): 219–234

McGuire RA, Ohsfeldt RL (1989b) Public choice analysis and the ratification of the constitution. In: Grofman B, Wittman (Hrsg) D (eds) The Federalist papers and the new institutionalism. Agathon, New York, pp 175–204

North DC (1981) Structure and change in economic history. Norton, New York

Oates W (1999) An essay on fiscal federalism. J Econ Lit 37(3):1120–1149

Oates W (2005) Toward A second-generation theory of fiscal federalism. Int Tax Public Financ 12(4):349–373

Persson T, Tabellini G (2000) Political Economics – Explaining Economic Policy. Cambridge et al.: The MIT Press

Persson T, Tabellini G (2003) The economic effects of constitutions. The MIT Press, Cambridge, MA

Persson T, Roland G, Tabellini G (1997) Separation of powers and political accountability. Q J Econ 112:310–327

Persson T, Roland G, Tabellini G (2000) Comparative politics and public finance. J Polit Econ 108(6):1121–1161

Rawls J (1971) A theory of justice. Belknap, Cambridge

Robinson J, Torvik R (2008) Endogenous Presidentialism. available at: http://www.nber.org/papers/w14603

Rodden J (2003) Reviving leviathan: fiscal federalism and the growth of government. Int Organ 57:695–729

Tanzi V (2000) Some politically incorrect remarks on decentralization and public finance. In: Dethier J-J (ed) Governance, decentralization and reform in China, India and Russia. Kluwer, Boston, pp 47–63

Tiebout C (1956) A pure theory of local expenditures. J Polit Econ 64:416–424

Voigt S (1999) Explaining constitutional change – a positive economics approach. Elgar, Cheltenham

Voigt S (2011) Positive constitutional economics II – a survey of recent developments. Public Choice 146(1–2):205–256
Voigt S (2013) Veilonomics: on the use and utility of veils in constitutional political economy. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2227339
Voigt S, Blume L (2012) The economic effects of federalism and decentralization: a cross-country assessment. Public Choice 151:229–254
Wicksell K (1896) Finanztheoretische Untersuchungen. Jena, Fischer

Further Reading

Voigt S (1997) Positive constitutional economics – a survey. Public Choice 90:11–53
Voigt S (2011) Empirical constitutional economics: onward and upward? J Econ Behav Organ 80(2):319–330

Constructivism, Cultural Evolution, and Spontaneous Order

Régis Servant
PHARE, University Paris 1 Panthéon-Sorbonne, Paris, France

Definition

This essay describes two completely different approaches which have been distinguished by Friedrich A. Hayek, the "constructivist" one and the "evolutionary" one, to the problem of how to develop institutions appropriate for the achievement of a desirable society. It does it by detailing Hayek's analysis of the two approaches. First, the essay describes the "constructivist" contention that only institutions deliberately adopted by certain competent persons are likely to achieve a desirable society. Then, it presents the "evolutionary" viewpoint defended by Hayek, according to which a lot of beneficial institutions can be discovered only through spontaneous, undesigned growth.

Constructivism

Hayek uses the term "constructivist rationalists" to designate a large and diverse group of scholars which includes Bacon, Descartes, Hobbes, Leibniz, Spinoza, Voltaire, Rousseau, Bentham, Austin, Hegel, Marx, Comte, Saint-Simon, and the American Institutionalists. These scholars, Hayek contends, are characterized by an arrogant overestimation of man's intellectual capacity to achieve a desirable society. They believe that people are, or at least can become, intelligent enough to make and remake the institutions of their society as they please so as to render those institutions better suited to their wishes.

For instance, the institutionalist Mitchell (1925) describes the organized group of pecuniary institutions which makes up our economic system as "marvelously flexible" and feels confident that the progress of statistical science will increasingly enable economists to reshape that flexible system, to make it better fitted to our needs. The development of conscious experiments on group behavior and the statistical analysis of their results, may, Mitchell contends, enable humanity to guide the evolution of its institutions more wisely, through a better awareness of the relation between institutions and consequences, "to convert society's blind fumbling for happiness into an intelligent process of experimentation" (1925, p. 8).

Hayek adds, more specifically, that, according to "constructivist rationalists," only institutions deliberately adopted by certain competent persons are likely to achieve a desirable society and that spontaneously grown institutions whose benefits are not clearly understood ought to be rejected. Applied to the field of law, for example, such a view means that, to be desirable, a system of law ought to emanate from the deliberate enactments of expert legislators capable of knowing beforehand what laws are better for the community. Thus, under a "constructivist regime," the development and alteration of the legal system, and of the whole framework of institutions more generally, would depend exclusively on rational foresight and understanding.

Hayek emphasizes one crucial consequence of such a view. From a political and legal standpoint, the logic of constructivism implies that people ought not to be authorized to introduce new institutions outside those deemed desirable by the ruling experts. Indeed, if one believes in the intellectual capacity of certain experts to know beforehand what institutions are better, then one is

naturally led to deny the rest of the people the legal right to try alternative institutions, on the ground that the trials would be useless anyway – a waste of time and energy. Constructivism thus tends to imply *a suppression of freedom*, or, in Hayek's words, "monopolistic power to experiment in a particular field – power which brooks no alternative and is in its essence based on a claim to the possession of superior wisdom" (1958, p. 242). In such a context, the only chance for an outsider to introduce alternative institutions would be to succeed in convincing the chosen experts that her proposal would prove superior. But, in any case, Hayek contends, it would rest exclusively with certain intelligent persons to decide, after rational reflection and argumentation, which institutions are worthy of adoption and which are not. The spontaneous growth of institutions, on the other hand, would not be tolerated, as Hayek describes in *The Constitution of Liberty*.

> It is worth our while to consider for a moment what would happen if only what was agreed to be the best available knowledge were to be used in all action. If all attempts that seemed wasteful in the light of generally accepted knowledge were prohibited and only such questions asked, or such experiments tried, as seemed significant in the light of ruling opinion, mankind might well reach a point where its knowledge enabled it to predict the consequences of all conventional actions and to avoid all disappointment or failure. Man would then seem to have subjected his surroundings to his reason, for he would attempt only those things which were totally predictable in their results. (1960, pp. 37–38)

Spontaneous Order

Hayek argues that the history of mankind repudiates the constructivist view. Certain institutions – such as language, writing, the family, money, the price system, and, more generally, the institutions of the market, as well as many other rules of morality and of law – have proven beneficial by assisting people to collaborate effectively, reduce conflicts, achieve a more economical utilization of resources, increase the level of general wealth, etc. *Yet these institutions did not emanate from a deliberate intention of certain experts capable of knowing beforehand that they would produce such general benefits*. Hayek speaks of "spontaneous growth" or "spontaneous order" to characterize that process by which a lot of beneficial institutions developed without people understanding clearly in what way they benefited their community. Scholars such as Mandeville, Vico, Hume, Burke, Smith, Ferguson, Savigny, Menger, and Maine are praised by Hayek for their "evolutionary" approach which emphasizes the indispensability of such undesigned, blindly growing institutions in the achievement of a desirable society. For instance, Menger (1883) argues that "*institutions which serve the common welfare and which are most important for its advancement can come into being without a common will directed towards establishing them*" (1883, p. 163).

One could retort to Menger, to Hayek, and to "evolutionary rationalists" more generally, that, although certain institutions which have grown spontaneously and whose benefits are not clearly understood may indeed benefit humanity, deliberately adopted institutions ought nevertheless to be exclusively preferred, for they can do even better. Yet, as explained below, Hayek maintains the exact opposite view in his theory of cultural evolution.

Cultural Evolution

Hayek presents a counterfactual history of how societies would have looked if, hitherto, the development and alteration of social institutions had been guided exclusively – or even mainly – by certain people's capacities for understanding and foresight. Under such circumstances, Hayek argues, the institutions available to people would actually have been very primitive. Not only we would be nowadays "much poorer" and "less wise," writes Hayek, "but we would also be less gentle, less moral; in fact we would still have brutally to fight each other for our very lives" (1968, p. 243).

According to Hayek, people discovered the vast majority of their beneficial institutions precisely because they did *not* endorse a constructivist rationalism. That is, they tolerated the growth

of certain institutions such as language, writing, the family, money, the price system, etc., despite the fact that the desirability, for the community, of such institutions was not, at the time, predictable or intelligible.

The chief explanatory example of Hayek's theory of cultural evolution relates to the development of the institutions of the market. People long lived within small food-sharing groups held together by highly collectivist institutions, Hayek explains, until a group of pioneers (period one, *t*) stumbled upon new rules of conduct, rules of the law of property, tort, contract, etc., which are at the root of the development of a market system, yet without intending thereby to achieve a more economical utilization of resources nor foreseeing the subsequent immense increase of peoples' general wealth. Indeed, those pioneers, says Hayek, "simply started some practices advantageous to them, which then did prove beneficial to the group in which they prevailed" (1979, p. 161).

Next came a second period (*t*+1) when the groups who had adopted the institutions of the market became more prosperous and more prolific than others and gradually displaced (or were imitated by) them. And only long after they grew up (*t*+2) did these superior institutions begin to be understood by a few professional economists. At the time of their development (*t*), their benefits as regards the utilization of resources and the level of general wealth were not, and could not have been, foreseen by anyone. This is what Hayek explained, for instance, in his Morrell Memorial Lecture on toleration published in 1987.

> And to me the proof of this is that even now hardly anybody yet understands what the advantages of private property and the market society are. Man not only did not know what the advantages would be when he introduced it, he still does not understand it fully today although he owes to it (and now I am coming to one of my chief points) the possibility of multiplying the numbers of mankind by roughly 200 times. (1987, p. 40)

Hayek infers that if people had endorsed a constructivist rationalism, they would have rejected the institutions of the market from the very beginning (at *t*), which would have been a wrong choice. Yet as it was, deviations from constructivism enabled some groups to be selected by *cultural evolution* for having adopted a superior sort of institutions. Indeed, under cultural evolution, Hayek insists, institutions are ultimately chosen by an impersonal group selection where the superior practices become known only *afterward* on the basis of what turns out to be more successful, *not beforehand* on the basis of what a body of alleged experts considers worthy of adoption. The following passage presents Hayek's evolutionary viewpoint very clearly.

> Man never deliberately created the institutions of private property or the family, or understood why he accepted the moral practices that they entail. The morals of property and the family were spread, and came to dominate a large part of the world, not because those who accepted them were able rationally to convince others that they were correct, and certainly not because they themselves liked them, but because those groups who by accident did accept them prospered and multiplied more than others. Thus we owe our morals not to our intelligence but to the fact that some groups uncomprehendingly, and indeed unwillingly (for mankind has been civilized against its wishes), accepted certain rules of conduct – the rules of private property, of honesty, of the family – and thus enabled the groups practising them to prosper, multiply, and gradually to displace other groups. (1986, p. 689)

Political Consequence

That spontaneous growth may enable people to discover institutions which are better than what deliberate design can ever achieve does not imply, however, that people ought to abstain from employing their capacity of understanding and foresight – as Hayek indeed recognizes. The prescriptive conclusion of the evolutionary approach is, rather, that, whatever their degree of expertise, people ought to acknowledge that the institutions they are capable of deliberately designing may not be the best ones and that alternative institutions whose benefits cannot fully be grasped may prove to be a superior means to success. Therefore, under an "evolutionary regime" à la Hayek, by contrast to a "constructivist" one, each person would be granted the legal right not to be constructivist. That is, the law would, on the one hand, authorize each person to accept certain institutions whose benefits are not clearly understood

and, on the other hand, authorize them to brave the opinion of experts, to try new institutions without being required to rationally demonstrate to any official authority the appropriateness of her proposal to achieve a desirable society.

Future Directions

Many scholars have argued that the "evolutionary rationalism" defended by Hayek implies an apology for freedom of experiment as against government monopolies (see, for instance, Arnold (1980), Gissurarson (1987), Steele (1994), Petroni (1995), Macedo (1999), Ebenstein (2001), Steele (2002), Salle (2003), and Servant (2014, 2018)). Yet, as regards some fields such as the development and alteration of the constitution, the law, a minimum social safety net system, etc., Hayek does not advocate freedom of experiment but, on the contrary, a monopolistic power of control. That a complete abrogation of the State's exclusive power to experiment is *not* deemed desirable by Hayek signifies that, despite his evolutionary approach, one could recognize *limits* to the appropriate domain of spontaneous order. How and where, then, to locate the boundary between monopolistic and competitively developed institutions? Moreover, an important problem arises from Hayek's acknowledgment that people who adopt spontaneously grown institutions very often do not understand in what way they benefit their community. In particular as regards institutions which, besides their benefits, also have great costs and may at times involve great disappointments, if people cannot see why they ought to accept those beneficial institutions, the act of accepting them might cause not only an increase but, at the same time, a *decrease* of welfare. In such a context, an area for further research could concern the question of what meaning exactly ought to be attached to the term "better institutions" from an evolutionary perspective.

Cross-References

- ▶ Austrian School of Economics
- ▶ Bounded Rationality
- ▶ Customary Law
- ▶ Hayek, Friedrich August von
- ▶ Institutional Economics
- ▶ Liberty
- ▶ Private Property: Origins
- ▶ Property Rights: Limits and Enhancements
- ▶ Rule of Law

C

References

Arnold R (1980) Hayek and institutional evolution. J Libertarian Stud 4(4):341–352

Ebenstein A (2001) Liberty and law. In: Ebenstein A Friedrich Hayek A biography. The University of Chicago Press, Chicago, pp 223–226

Gissurarson H (1987) Three liberal criticisms of traditionalism. In: Gissurarson H Hayek's conservative liberalism. Garland, New York, pp 126–132

Hayek F (1958) Freedom, reason, and tradition. Ethics 68(4):229–245

Hayek F (1960) The constitution of liberty. University of Chicago press, Chicago

Hayek F (1968) Speech on the 70th birthday of Leonard Reed. In: What's past is prologue: a commemorative evening to the foundation for economic education on the occasion of Leonard Read's seventieth birthday. Foundation for Economic Education, New York, pp 37–43

Hayek F (1979) Law, legislation and liberty, vol. 3: the political order of a free people. Routledge & Kegan Paul, London

Hayek F (1986) Cultural evolution: the presumption of reason. World & I 1(2):683–690

Hayek F (1987) Individual and collective aims. In: Mendus S, Edwards D (eds) On toleration. Clarendon Press, Oxford, pp 35–47

Macedo S (1999) Hayek's liberal legacy. Cato J 19(2): 289–300

Menger C (1883) Untersuchungen über die Methode der Sozialwissenschaften, und der politischen Oekonomie insbesondere. Verlag von Duncker & Humblot, Leipzig

Mitchell W (1925) Quantitative analysis in economic theory. Am Econ Rev 15(1):1–12

Petroni A (1995) What is right with Hayek's ethical theory. Rev Eur Sci Sociales 33(100):89–126

Salle C (2003) Fin de l'Histoire et légitimité du droit dans l'oeuvre de F. A von Hayek. Rev Fr Sci Polit 53(1): 127–166

Servant R (2014) Libéralisme, socialisme et État providence : la théorie hayékienne de l'évolution culturelle est-elle cohérente ? Rev économique 65:373–390

Servant R (2018) Spontaneous growth, use of reason, and constitutional design: Is F. A. Hayek's social thought consistent? J Hist Econ Thought, forthcoming

Steele G (1994) On the internal consistency of Hayek's evolutionary oriented constitutional economics: a comment. J Économistes Études Hum 5(1):157–164

Steele G (2002) Hayek's liberalism and its origins: his idea of spontaneous order and the Scottish enlightenment. Christinia Petsoulas. Q J Austrian Econ 5(1):93–95

Further Reading

Boudreaux D (2014) Legislation is distinct from law. In: Boudreaux D The essential Hayek. Fraser Institute, Canada, pp 32–38
Buchanan J (1977) Law and the invisible hand. In: Buchanan J Freedom in constitutional contract: perspectives of a political economist. Texas A&M University Press, College Station, pp 25–39
Diamond A (1980) F. A. Hayek on constructivism and ethics. J Libertarian Stud 4(4):353–365
Hayek F (1946) Individualism: true and false. Hodges, Figgis & Co. Ltd., Dublin
Hayek F (1965) Kinds of rationalism. Econ Stud Q 15(2):1–12
Hayek F (1966) Lecture on a master mind: Dr. Bernard Mandeville. Proc Br Acad 52:125–141
Hayek F (1967) Notes on the evolution of systems of rules of conduct. In: Hayek F Studies in philosophy, politics and economics. University of Chicago press, Chicago, pp 66–81
Hayek F (1973) Law, legislation and liberty, vol. 1: rules and order. Routledge & Kegan Paul, London
Hayek F (1988) The fatal conceit: the errors of socialism. Routledge, London
Kirzner I (1990) Knowledge problems and their solutions: some relevant distinctions. Cult Dyn 3(1):32–48
Moroni S (2014) Two different theories of two distinct spontaneous phenomena: orders of actions and evolution of institutions in Hayek. Cosmos + Taxis 1(2):9–23
Nadeau R (2016) Cultural evolution, group selection and methodological individualism: a plea for Hayek. Cosmos + Taxis 3(2):9–22
O'Driscoll G (2015) Hayek and the scots on liberty. J Priv Enterp 30(2):1–19
Smith V (2008) Rationality in economics: constructivist and ecological forms. Cambridge University Press, Cambridge
Smith C (2014) Hayek and spontaneous order. In: Garrison R, Barry N (eds) Elgar companion to Hayekian economics. Edward Elgar Publishing Limited, Cheltenham, pp 224–245

Consumer Bias

Sophie Bienenstock
Economix, Université Paris Nanterre, Paris, France

Abstract

Since the founding work of Simon (1955) and Kahneman and Tversky (1974, 1986), it is recognized that agents, namely consumers, are endowed with a bounded rationality. Consumer biases cover a wide range of behaviors, such as quality misperception, status quo bias, projection bias, inertia, and can have various consequences on the market equilibrium. The aftermaths of consumer misperception depend on the type of bias one considers, as well as on the market structure. This entry presents a typology of consumer biases and mentions possible consequences on the market outcome. Policy recommendations to fight against consumer biases, as well as the main counterarguments put forward by libertarians, will finally be discussed.

Definition

Bounded rationality refers to the fact that agents depart in a systematic way from the perfect rationality assumption which prevails in the economic literature (for a more general approach of the topic, see the entry on "▶ Bounded Rationality"). As summarized by Herbert Simon, the ambitious aim of behavioral models of rational choice is to "replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and the computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist" (Simon 1955, p. 99).

While the issue of bounded rationality is not constrained to the case of consumers (see for instance Rabin 2002 and Kahneman 2011). the latter are particularly prone to various kinds of cognitive biases, because of the context as well as the intrinsic particularities of consumption decisions (Korobkin 2003). Consumers are bound to make numerous and complex decisions, which leads them to use simplifying heuristics. Moreover, they face sophisticated and rational firms, who are likely to enhance or even exploit their weaknesses. Hence, consumer behavior is a particularly prosperous field for behavioral biases. Recent developments in Law and Economics have therefore brought the issue of consumer decision-making back to the center stage. The main

concern is to understand consumer behavior in order to suggest relevant responses in terms of public policy. While the existence of consumer biases is unanimously recognized, the issue of whether and how the regulator should intervene remains debated.

Theoretical Work on Consumer Bias

The main issues tackled in the literature dedicated to consumer biases are twofold:

- First, one needs to describe and understand the consequences of consumers' cognitive limitations. What impact do biases have on consumer choice, and consequently on competition and pricing? How does the market equilibrium change in the presence of consumer biases?
- Once this first step has been addressed, one can tackle the issue of fighting against consumer biases. The natural question that comes to mind is whether inefficiencies linked to consumer biases can be reduced by increasing competition. Will biased consumer behavior be overcome through learning or education?

Such concerns have become central in the economic literature, to the point that some authors assert that "the rational firm-irrational consumer assumption has become the norm, and the question of what firms do to exploit irrationality is often the primary focus" (Ellison 2006). Yet, describing consumer biases, as well as their consequences on the market, is not an easy task. Behind the notion of consumer biases lies a great diversity of behaviors. Consumer biases are numerous and do not refer to a unique cognitive phenomenon. Hence, classifying consumer biases is an essential step towards understanding their economic consequences.

Classifying Consumer Biases Although classifying the numerous biases is an unrelenting and complex task, Huck and Zhou offer a simple typology. The authors identify three dimensions along which consumer choices might be biased (Huck and Zhou 2011).

- First, the willingness to pay bias describes a situation in which agents pay too much for a given quantity of a good. For instance, according to the reference point effect, an agent's willingness to pay depends on a reference point (the status quo, past experience, other products, expectations, etc.). The reference point might lead to an irrational increase in the agent's willingness to pay. Willingness to pay bias can also be due to a misperception of future desired attributes. As in the famous study of DellaVigna and Malmendier (2006), consumers might believe that they will go to the gym more than they actually will. This misperception can be analyzed as overoptimism and willingness to pay bias. Similarly, the common knowledge according to which shopping on an empty stomach leads to overconsumption is another expression of such misperception (Loewenstein et al. 2003).
- Second, a search bias occurs when consumers do not choose the best-suited product because they do not search in a rational way. Consumers might for example be subject to inertia, which is to some extent an expression of the well-known status quo bias. More generally, inertia describes consumers who tend to buy in one of the first shop they see and/or to stick to the same seller in case of repeated purchase, although more advantageous offers might be available. Another form of search bias is price misperception. In the presence of complex price schemes (such as partitioned pricing, drip pricing, baiting, discounts) consumers are likely to misjudge the final price. In response to price misperception, firms might have incentives to adopt artificially complex pricing systems (Spiegler 2006).
- Finally, quality biases refer to any situation in which consumers purchase a quality not fit for their needs. The error can be due to a misperception of the intrinsic quality of the product, or to inaccurate anticipations about one's own needs and capacities to use a product. This observation leads us to the interesting dichotomy proposed by Köszegi (2014): false beliefs either relate to "the contract itself," or to "her own behavior given the contract" (p. 1104).

Modeling Consumer Biases Since consumer biases cover a large scope of behaviors, there is no general model of consumer decision-making in the presence of cognitive bias. Studying consumer biases requires building an ad hoc model, which depends both on the kind of cognitive flaw one wants to describe and on the market structure one considers. Therefore, theoretical articles dedicated to consumer biases generally focus on one specific type of irrationality in a given market structure. One should always keep in mind that the conclusions are necessarily context-dependent and that formulating a general theory of consumer bias is by essence very difficult.

Nonetheless, several papers make a significant contribution in the understanding of consumer biases. For instance, hyperbolic discounting is at the center of several papers by DellaVigna and Malmendier (2004, 2006). Time inconsistent preferences have been studied by Eliaz and Spiegler (2006), while the framing effect is analyzed by Piccione and Spiegler (2012).

The consequences of consumer biases depend on numerous parameters. Yet, authors generally come to the conclusion that consumer biases are detrimental to social welfare because they result in "behavioral market failures" (Bar-Gill 2011). The concept of "behavioral market failures" was coined by Oren Bar-Gill (2011) to describe the deficiencies of market mechanisms in the presence of boundedly rational consumers. Hence, the issue of whether and how one should intervene to constrain the aftermaths of consumer biases naturally comes to mind.

Policy Implications: Should the Regulator Intervene?

In the broad lines, two kinds of responses to consumer biaises are conceivable: soft paternalism on the one hand, and debiasing, on the other hand. While they both strive towards a common goal, they differ in the method they use.

Soft Paternalism The notion of soft paternalism, also known as asymmetric paternalism (Camerer et al. 2003) or libertarian paternalism (Sunstein and Thaler 2003), refers to any legal intervention aimed at protecting agents without encroaching on individual freedom. The concept has been thoroughly studied by Sunstein and Thaler (2008) in their book *Nudge: Improving Decisions About Health, Wealth and Happiness*. Soft paternalism claims to be both ostensibly paternalistic, in that it helps people make decisions that are in their best interest, and libertarian, in that it preserves freedom of choice.

Debiasing Debiasing consists in revealing errors to each agent, so that they can correct their behavior on their own (Jolls and Sunstein 2006). Debiasing differs from paternalism insofar as consumers are fully aware of the process they undergo. The ultimate objective is to give agents the capabilities to act by themselves. Sunstein and Thaler (2008) clearly sum up the concept of debiasing in what they call "RECAP Policies" (Record, Evaluate, Compare, Alternative Prices). While soft paternalism relies on manipulation, debiasing rests on increased transparency. Both raise vivid criticisms from libertarians.

The Libertarian Criticisms The first main criticism, addressed mainly to soft paternalism, concerns the complexity of carrying out a welfare analysis in the presence of consumer bias. According to the libertarian view, the regulator does not have the relevant information to determine the agents' true preferences in the presence of changing utility functions or inconsistent choices. In this line of thought, Saint-Paul (2011) claims that "it is impossible, in fact, to establish such a result, for one needs a criterion for comparing alternative utility functions; that is, one would have to impose some 'meta-utility function' in order to tell us that a given utility function is better than another" (p. 87). Yet, such a meta-utility function does not exist, which renders any assessment on welfare impossible in the presence of changing preferences.

The second main argument can be summed up as the fear of a "slippery slope," according to which there is a natural tendency to go towards

more paternalism. This slippery slope, which would lead to strong paternalist measures and to the denial of individual freedom, has been put forward by Whitman and Rizzo (2007, 2009).

Beyond the slippery slope argument, also lies the idea that too much regulation would inhibit learning, and ultimately be counterproductive. A systematic intervention to guide citizens towards what is considered to be a good decision would remove all opportunities to make errors. In this line of thought, Klick and Mitchell (2006) allege that regulation leads to a vicious circle of more regulation, which ultimately increases biases. In this approach, cognitive biases are considered to be endogenous, insofar as they depend on the existing regulation.

The last main argument put forward by libertarians is that legal interventions to counter the effects of cognitive biases are useless, since the market remains efficient even if agents are not perfectly rational. For instance, Sugden (2008) contends that the market is an efficient way of allocating resources even if consumers exhibit inconsistent preferences. Sugden's key argument lies in the fact that firms always have incentives to cater to consumer demand, in spite of potentially inconsistent preferences. The idea that the market is the best response to consumer bias has also been suggested by Bebchuk and Posner (2006).

Conclusion

While the ubiquity of consumer biases cannot be denied, the appropriate response remains debated. In various countries, consumer policy is slowly starting to take into consideration the presence of consumer biases, for instance, by using the framing effect in order to steer consumers towards healthy foods. Yet, such measures remain sparse. One of the reasons is that no general policy to fight consumer bias can be enacted, since every situation requires an ad hoc analysis. In spite of the work that remains to be done, we believe that several powerful and simple tools can be used to protect consumers against their own flaws.

Cross-References

- ▶ Behavioral Law and Economics
- ▶ Bounded Rationality
- ▶ Consumption
- ▶ Contract, Freedom of
- ▶ Endowment Effect
- ▶ Harmonization: Consumer Protection
- ▶ Naïve Consumers: Contract Economics
- ▶ Rationality

References

Bar-Gill O (2011) Competition and consumer protection: a behavioral economics account. Law & Economics Research paper series, working-paper no 11–42

Bebchuk L, Posner R (2006) One-sided contracts in competitive consumer markets. Mich Law Rev 104: 827–836

Camerer C, Issacharoff S, Loewenstein G, O'Donoghue T, Rabin M (2003) Regulation for conservatives: behavioral economics and the case for "asymmetric paternalism". Uni Pennsylvannia Law Rev 3(151):1211–1254

DellaVigna S, Malmendier U (2004) Contract design and self-control: theory and evidence. Q J Econ 119(2): 353–402

DellaVigna S, Malmendier U (2006) Paying not to go to the gym. Am Econ Rev 96(3):694–719

Eliaz K, Spiegler R (2006) Contracting with diversely naive agents. Rev Econ Stud 73(3):689–714

Ellison G (2006) Bounded rationality in industrial organization. In: Blundell R, Newey WK & Persson T (eds), 'Advances in economics and econometrics: theory and applications', Vol. 2 of Ninth World Congress, Cambrige university Press, pp. 142–219. http://economics.mit.edu/files/904

Huck S, Zhou J (2011) Consumer behavioral biases in competition: a survey, Final Report OFT1324, Office of Fair Trading

Jolls C, Sunstein C (2006) Debiasing through law. J Leg Stud 35(1):199–242

Kahneman D (2011) Thinking fast and slow. Farrar, Staus and Giroux, New York

Kahneman D, Tversky A (1974) Judgment under uncertainty: heuristics and biases. Science 185(4157): 1124–1131

Kahneman D, Tversky A (1986) Rational choice and the framing of decisions. J Bus 59(4):S251–S278

Klick J, Mitchell G (2006) Government regulation of irrationality: moral and cognitive hazards. Minnesota Law Rev 90:1620–1663

Korobkin R (2003) Bounded rationality, standard form contracts, and unconscionability. Uni Chicago Law Rev 70(4):1203–1295

Köszegi B (2014) Behavioral contract theory. J Econ Lit 52(4):1075–1118
Loewenstein G, O'Donoghue T, Rabin M (2003) Projection bias in predicting future utility. Q J Econ 118(4):1209–1248
Piccione M, Spiegler R (2012) Price competition under limited comparability. Q J Econ 127:97–135
Rabin M (2002) A perspective in psychology and economics. Eur Econ Rev 46:657–685
Saint-Paul G (2011) The tyranny of utility: behavioral social science and the rise of paternalism. Princeton University Press, Princeton
Simon H (1955) A behavioral model of rational choice. Q J Econ 69(1):99–118
Spiegler R (2006) Competing over agents with boundedly rational expectations. Theor Econ 1(2): 207–231
Sugden R (2008) Why incoherent preferences do not justify paternalism. Constit Polit Econ 19: 226–248
Sunstein C, Thaler R (2003) Libertarian paternalism. Am Econ Rev 2(93):175–179
Sunstein C, Thaler R (2008) Nudge: improving decisions about health, wealth and happiness. Yale University Press, New Heaven, CT, USA
Whitman DG, Rizzo M (2007) Paternalist slopes', NYU. J Law Econ 2:411–443
Whitman DG, Rizzo M (2009) The knowledge problem of new paternalism. Brigham Young Uni Law Rev 4:904–968. http://digitalcommons.law.byu.edu/lawreview/vol2009/iss4/4/

Further Reading

Bernheim D, Rangel A (2009) Beyond revealed preferences: choice-theoretic foundations for behavioral welfare economics. Q J Econ 124(1):51–104
Rabin M (2002) A perspective in psychology and economics. Eur Econ Rev 46:657–685
Rachlinski J (2003) The uncertain psychological case for paternalism. Northwest Univ Law Rev 97(3):1165–1225
Spiegler R (2011) Bounded rationality in industrial organization. Oxford University Press, Oxford

Consumer Law

▶ Harmonization: Consumer Protection

Consumer Legislation

▶ Harmonization: Consumer Protection

Consumer Policy

▶ Harmonization: Consumer Protection

Consumption

Fav Tsoin Lai
National Chi Nan University, Puli, Taiwan

Abstract

Consumption involves consumer behavior of how people make choice in selecting to enjoy bundle with specific quantities of one or more goods and services within a period. Economists use an often-used word "demand" to term peoples' choices when consuming a good, and model the choices beginning with an account of consumer preferences over commodity bundles. Goods are categorized as normal or inferior by the responses of consumers' demands to the income changes, and economists break the effect of price changes on demands into two parts, namely, substitution and income effects. People have their own time preference over the consumptions in different periods. Two cases are illustrated: one is that current consumptions are the substitutes for future consumptions, and the other is that the pleasure of current consumption is enhanced by the past consumptions. Decisions people make regarding consumptions not only occur in the economic situations with certainty but also with uncertainty. Most of economists nail uncertainty down as risk, and it refers to situations, in which consumers can list all possible outcomes and subjectively know the likelihood of each occurring. The ways that people rank plans of consumption under uncertainty are similar to the ones that people have preferences over consumption bundles under certainty. An individual may be not only concerned with his/her own self, but also be connected to the selves of the others, and hence his/her demand for some goods could depend on the

demands on the part of other consumers. In the case of positive network externalities, the interdependence preferences boost the demand for a good or service, but in the case of negative ones they reduce the consumption.

Definition

The using up of goods or services by individuals' choice.

Consumption can be understood as the using up of goods or services by individuals' choice. It involves consumer behavior of how people make choices in selecting to enjoy bundle with specific quantities of one or more goods and services within a period. Economists use an often-used word "demand" to term peoples' choices when consuming a good. The individual's demand for a good can be independent of another person's and just associated with his/her own tastes and income and the prices of goods. However, for some goods, one person's demands are related to the demand of other people. While discussing consumers' demands for goods, most of microeconomics textbooks start from and focus on the choices that are not associated with social interactions.

The scholars of economics develop a practical way to describe the reasons why an individual makes a specific good or service not others; they clarify what is affordable, describe the assumptions on which preferences are based, present how consumers make consumption decisions, and analyze the characteristics of demands. All the combinations of goods that the consumer may choose are referred to as consumption bundles or market bundles. Each bundle is a list with specific quantities of goods and services. Consumers can afford the bundles that do not cost more than his/her income, and the collection of such affordable bundles is referred to as budget set. The relative price (price ratio) can be a measure that fathoms the opportunity cost of consuming one good in terms of the other good. In a well-organized market, it is a kind of objective exchange rate in which most of the people will trade one good for another.

Consumer Behavior

Some economists believe that their impositions on the ordering of consumer preferences hold for most people in most situations. The first one is completeness. Consumers are able to compare any two bundles, rank them, and hence make a choice between them. The second one is that consumer preferences are transitive. It says that if bundle A is preferred to bundle B and bundle B is preferred to C, then it must be the case that A is preferred to C. Transitivity is a hypothesis about people's choice behavior and is normally regarded as necessary for consumer consistency. People, who violate this condition, have circular preference that they cannot have the same attitude toward a thing and are often contradictory when dealing with the logically connected events. Usually, economists believe that these two characteristics, completeness and transitivity, are the most important parts of consumers' rationality. (For more details about the assumptions about consumers' preferences, refer to Varian (1992), Mas-Colell et al. (1995), and Pindyck and Rubinfeld (2012a)).

As long as the goods do not satiate the consumers, they always prefer more of any goods as opposed to less. In this case, consumer preferences are referred as monotonic by economists. There could be further assumptions regarding consumer preferences, which make economists' practices proceed more smoothly. For example, well-behaved preference has the traits: averages are preferred to extremes, and the ordering of consumer preferences is continuous. The former trait seems to be a usual observation of ordinary peoples' preference, and the latter indicates that a small change in the combination of goods will not cause a significant disturbance in their ordering of preferences. All these assumptions provide the classic, preference-based approach to illustrate how a person's actual consumption is determined.

After recognizing that consumer's choice over bundles is only related to the rank of the satisfactions between two bundles and not the metrics of satisfactions, economists construct a utility function in such a way that the number assigned to the more-preferred bundle is larger than the one

assigned to the less-preferred bundle. The utility function presents the order of preferences over the bundles, not how much one bundle is preferred to the other, so that the magnitude of the difference between the utilities of the two bundles does not matter. What matters is the order of preferences, not the strength of people's preference. When the quantity of good changes by a very small amount, the number assigned will be changed; the ratio of the change in utility to the quantity change of the good is referred as marginal utility with respect to it. Since the magnitude of utility is for ordinal function, the marginal utility does not have behavior content. All the bundles that have the same utility constitute a set in the commodity space that is referred as indifference curve by economists. Any bundle in the indifference curve has the same utility and hence can be used to establish a measure that is called marginal rate of substitution – the maximum amount of one good that a person is willing to give up to obtain an additional unit of the other good. It is a consumer's subjective exchange rate at which a consumer is willing to trade one good for the other.

The preference of a consumer is subjective; everyone has his/her own taste and ranks his/her preference over consumption bundles. The diversity of personal tastes gives rise to a range of preference orderings that are very different from each other. Economists insist that no metric exists enabling one to make interpersonal comparison of well-being. The difference between the utility a person assigns to a bundle and the utility the other person assigns to it tells nothing about what the difference between their satisfactions is. An individual with subjective preference can compare different groups of items available for purchase and rank all of them, but there are many restrictions on the quantities of goods they can buy. Limited real incomes are the foremost constraint that forces consumers to choose among alternatives. Consumers consider not only nominal income but also the prices of goods because the amount of money they spend on the goods should be no more than the total amount they can spend. Higher income and lower prices make consumers' budget set larger and allow them to afford more. Given their rationality and limited real income, consumers make optimal choice over affordable set; that is, they choose the best to consume and maximize their utility. Consumers' consumption or demand is a function of prices and their incomes. When a consumer makes a choice involving all kinds of goods, his/her subjective exchange rate will be equal to the objective exchange rate. This is a marginal condition for consumers' consumption choice – the marginal rate of substitution is equal to the ratio of prices.

Consumers are concerned with the changes in their incomes and the prices of goods, because these changes in economic environment affect what they can afford. When their budget sets are changed, consumers will adjust their choice to make themselves as good as possible. Economists examine the impacts of these changes on consumers' demand independently. In the case where the prices remain unchanged, the goods are described as normal: consumers want to buy more of the goods as their incomes increase. Some economists define normal goods further according to consumers' demand responses to the income changes. If the demand for a good goes up by a greater proportion than income (the demand elasticity of income is greater than one), it is called luxury good, and if it goes up by a smaller proportion than income, it is called necessary good (the demand elasticity of income is less than one). Alternatively, such a good is called inferior good: an increase of income results in a reduction in the consumption of the good. As a person's purchasing power increases, his/her satisfaction from the consumption should not decrease. Thus, for a well-behaved preference, it is impossible for all goods to be inferior.

Price adjustments for a good not only change the market exchange rates between it and the other goods but also affect consumers' real purchasing power. Economists break up the effect of price changes on consumers' demands for goods into two parts, namely, substitution and income effects. A good becomes cheaper as its price falls, and consumers will tend to buy more of it, but the other goods become relatively more expensive now, and consumers will be inclined to buy less of them. The change in price allows consumers to substitute cheaper goods for the

relatively expensive goods; economists describe this kind of response as the substitution effect, and it always moves in the opposite directions to the price changes. A fall in the price of a good not only allows consumers to buy the original choice but also enables them to afford extra amount of goods in their money income. This is equivalent to the movement that occurs when purchasing power goes up while the relative prices remain constant. Consumers would adjust their demand following the changes in their real income; this response is called the income effect. If the good is normal, the increases of purchasing power that is due to the fall in its price will lead to an increase in its demand, and then the direction of income effect will be opposite to the movement of its price. However, if the good is inferior, the direction of income effect will be the same as the movement in price. There are no definite signs for income effect. For a good, the impact of its price changes on its demand is related to how consumers' choices in regard to consumption respond to income change. If it is a normal good, income effect should reinforce the substitution effect, and its demand for the good must increase when its price decreases. However if it is an inferior good, the income effect has the movement that is opposite to the one for substitution effect, and the demand for the good does not always increase when its price falls. For an inferior good, if income effect is strong enough to dominate substitution effect, a fall in its price trumps consumers' demands for it. In considering this event, economists call such a good a Giffen good. The income effect is usually small compared to the substitution effect because most of the expenditures on individual good make up a small part of consumers' budgets. In addition large income effects are often related to normal good, seldom to inferior good. Thus, in the real world, Giffen goods are rarely to be encountered, and instead it is ubiquitous for people to consume more of a good as its price falls (Pindyck and Rubinfeld (2012a) Microeconomics, Pearson, Taipei).

For many goods, the demand for one good is often associated with the consumptions and prices of other goods. For the goods like rice and wheat that have some similar characteristics, when one good gets more expensive, the consumer switches to consume the other good; the consumer substitutes away from the more expensive good to the cheaper one. Two goods are substitutes if an increase in the price of one leads to an increase in the quantity demanded for the other. To some consumers, the pleasure of consuming a cup of coffee can be enhanced when they can consume with one or two spoons of sugar. Sugar and coffee are consumed together and complement each other. When sugar gets more expensive, the consumer not only consumes less sugar but also consumes less coffee; two goods are complements if an increase in the price of one leads to a decrease in the quantity demanded for the other.

Intertemporal Choice

In their life span, people have income streams that do not remain constant and have their own time preference over the consumptions in different periods, and hence their choices of consumptions are involved in saving and consuming over time. The returns of hoarding are the relative prices of consumption between periods and are equal to the principal of any amounts saved plus interest rate, and with income streams they form consumers' budget constraints. How much a consumer is willing to substitute consumption today for consumption tomorrow depends on his/her time preferences that are constituted by his/her particular patterns of consumptions over time. Time preference is a main factor that determines the intertemporal marginal rate of substitution at which a consumer is just on the margin of being willing to substitute current consumption for future consumption. Most economists take rates of time preference as given or exogenous for convenience in dealing with the subjects they focus on, but some economists think they could be changed by consumers' choices over time. (Uzawa (1968), Lucas and Stokey (1984), and Epstein (1987) consider the possibility that the rate of time preference ultimately depends on consumption flows. Becker and Mulligan (1997) postulate a model, in which a consumer can make

an effort to reduce the discount on future utilities.) A consumer who doesn't care whether he/she consumed today or tomorrow is such a patient person that there is no time discount between his/her consumptions in different periods; future consumptions can be perfectly substituted for current consumptions. The people, who value current consumption more than future consumption, have time impatience with delaying consuming a certain bundle of goods or services, there is a time discount against the values of their future consumptions. As what the consumers' choices in a static setup are, there exist substitutions between the consumptions at different periods, and it is also possible that consumptions can be complemented for each other.

If there is a complementary between the consumptions of a particular good at different times, the pleasure of current consumption is enhanced by the past consumption. Goods like cigarette and heroin are potentially addictive, because for a large number of consumers who consume these goods, their past consumption complements current consumptions (Becker and Murphy 1988). People that become addicted into smoking cigarettes will increase their current consumption if there is an increase in their past consumption. Because of the complementary between the consumptions today and that of tomorrow, a price increase in the price of tomorrow consumption will lead to a decrease in the consumption today. For example, an addicted smoker that anticipates there will be a tax on smoking may reduce his/her current consumption of cigarettes. The choices in regard to consumptions over time are very similar to the choices of consumptions over several goods within one period; there exists a complementary between the current and past consumptions. This can also be seen by the observation that environmental cues affect consumer behavior. The smell of cookies being baked, sound of ice cube falling into a whisky tumbler, and sight of a pack of cigarettes could be the cues for consuming the goods associated with them (Laibson 2001). Cigarette addicts feel a keen desire for nicotine when they see smoking cues, like an open box of cigarettes. Addict formation effects are triggered and halted by the occurrence and nonappearance of cues that have been related with the past consumption of addict forming goods.

Uncertainty

People's decisions makings regarding consumptions discussed in the previous paragraphs have been implicitly assumed to occur in the economic situations with certainty. However, there is considerable uncertainty involved in everyday life of people. Uncertainty can be interpreted broadly; however, most of the economists nail it down as risk, which refers to situations in which consumers can list all possible outcomes and subjectively know the likelihood of each occurring. Then, uncertainty is associated with the probability distribution that consists of a list of different outcomes and the subjective probability associated with each outcome. According to his/her experience and judgment, a person evaluates the possibility that an outcome will occur to form his/her subjective probability, and hence it is not necessary that this probability is the same as the rate of recurrence with which this outcome has actually occurred in the past. Since people subjectively gauge the likelihood that an outcome occurs, they may have different probability distributions over outcomes and different decision-making.

People can rank probability distributions as they can do the bundles of goods under certainty; that is, the way that consumers have preferences for various goods can be applied to the one in which consumers have preferences for different probability distributions. When an economic environment is characterized by uncertainty, a person will attach different values on a consumption bundle in different circumstances under which the consumption turns out to be reachable. Before they are sure which state of the world will occur, people have a plan that they will follow to have their consumption while facing different outcomes. An outcome of a random event can be interpreted as a state of nature, and what people actually consume is contingent on it. A plan for consumption under uncertainty constitutes of probabilities and outcomes, and it is a risky

alternative. Since each outcome gives an individual a level of satisfaction, he/she can evaluate every risky alternative according to his/her expectations and rank all risky alternatives. The ways that people rank plans of consumption under uncertainty are similar to the ones that people have preferences over consumption bundles under certainty. The difference between the consumer's choice under certainty and the one with uncertainty is the independent assumption for the outcomes under uncertainty. What an individual intends to consume in one state is independent of the choices he/she makes in other states, since the consumption choices in the different states of nature cannot coexist. Except this, the theory of consumers' choice under certainty can be applicable for analyzing people's choices among different risky alternatives. Not only can an individual's rationality, completeness, and transitivity be applied to his/her preference over risky alternatives, but also the other assumptions regarding the preferences over actual bundles are also applicable to them under a particular structure. Each outcome corresponds with consumption and can be assigned a number to represent the pleasure it brings up to a consumer. Then weighting the number by the probability of its occurrence and summing all the weighted values establish a consumer's expected utility of a risky alternative.

In the presence of uncertainty, there are two important measures that are used to describe a risky alternative, expected value and variability. Each outcome is associated with a payoff and a probability; the expected value of a risky alternative is the sum of payoffs being weighted by probabilities. The expected value is the payoff that an individual would expect on average, and hence it reveals the main propensity for the distribution of outcomes. The variability of a risky alternative is the spread of possible outcomes and tells the association with the riskiness of a random event. Most people love the risky alternative with high expected value and low variability. In addition, an individual's expected satisfaction about a risky alternative is related to but not necessarily proportional to its expected monetary values. People's attitudes toward risk determine their choices over risky alternatives and therefore their actual consumption. Economists categorize people's willingness to bear risk by identifying whether an individual ranks a certain income over an uncertain income with the same expected value. (This definition of consumers' preferences toward risk can be seen in Pindyck and Rubinfeld (2012b)). An individual who is risk averse prefers a certain income to a risky income with the same expected value; that is, he/she would rather have the expected value of his/her wealth rather own the lottery. A person who is risk neutral is indifferent between a certain income and an uncertain income with the same expected value. An individual who is risk loving prefers an uncertain income to a certain one, even if the expected value of uncertain income is less than that of certain income. However, some experiments to see the extent to which people's attitude toward risk fit those three propensities reveal that most people have the propensity to prefer avoiding losses over acquiring gains. This kind of multifaceted feelings about losses and gains is depicted by Kanehman and Tversky (Kahneman and Tversky 1979) as "loss aversion," which is often seen in inexperienced investors but not so often in experienced investors (List 2003).

Interdependence and Consumer Behavior

In the case where consumers have preferences independent of one another, their demands are only related to the prices of goods and their own incomes and tastes. However, an individual may be not only concerned with his/her own self but also be connected to the selves of the others, and hence his/her demand for some goods could depend on the demands on the part of other consumers. If the interdependence preferences boost the individual demand for a good or service, a positive network externality exists – the quantity of a good demanded by a usual consumer increases as the demands of other consumers increase (Leibenstein 1950). When people watch a game or play, they jointly consume the service it supplies in the presence of others, and this activity

becomes more worth having when they can be shared with a group of peers. Similarly, some social activities, like restaurant eating, attending a concert, talking about books, and chatting about the booming of stock markets, let a consumer experience more enjoyment the more that these activities are participated in by others (Becker 1991). Those have the common feature of a bandwagon effect – the longing for being fashionable, for owning a good because everyone has it, or to derive unconstrained pleasure from a fashion. The pleasure from a good is greater when many people want to consume it, perhaps because a person wishes to be in the same step with what is in fashion or because he/she is more certain that the food, writing, or performance has its quality when a restaurant, book, or theater is more popular. Moreover, because people take part in many activities with their families, coworkers, neighbors, and friends, interdependent preferences can be spread across and through several networks, which are channeled by those persons close to them.

Network externalities are sometimes negative. A specially designed sports car not only supplies a service of transportation but also brings prestige to its owner; it is a conspicuous good that is characterized by its exclusivity and provides its consumers with a signal of status, which is enhanced by material displays of wealth. Consumers purchase conspicuous good to satisfy their material needs and social needs as provided by the product. Conspicuous consumption is such an activity that consumers' indulgence in it is recognized by their peers and differentiates their consumption from that of the other groups. The desire to own the uniqueness and exclusivity of a conspicuous good motivates consumers to purchase it and produces a negative network externality, which is called snob effect by some economists (Leibenstein 1950). A branded watch has its instrumental function for its consumers, but what is more valuable is the prestige, status, and exclusivity resulting from the fact that not many people own one like it. When more of the others consume the conspicuous good, its status value declines. Thus an individual demand for a snob good decreases as the more people own it, and there is a negative feedback between individual consumption and aggregate consumption. An increase in the price of the good will reduce consumption but will enhance the status value of the good.

Since a conspicuous good signals the status of wealth, its price is much higher than the value of its instrumental function. Most of the conspicuous goods have branded names to show their exclusivity and to assure quality. The counterfeits of conspicuous goods do not have the same quality as genuine ones but have the appearance of being genuine. Thus, the producers of counterfeits separate the status and quality aspects of the product and thereby allow some consumers to purchase the former aspect even though they would not be willing to pay the high price of purchasing the two aspects together (Grossman and Shapiro 1988). People, who cannot afford the consumption of a genuine good, could consume its counterfeits that are in much lower price. Counterfeits become the available substitutes for the genuine good due to their price advantage, and hence the consumptions of counterfeits are getting widespread proportions in developing countries. The illegitimate producers may impose an externality on the owners of branded genuine goods by reducing the snob appeal of their possessions (Higgins and Rubin 1986). However, it has been documented that the consumption of some counterfeits can promote the sale of genuine goods. In the event that genuine goods can be differentiated from their counterfeits, the sale of counterfeits can enhance the exclusivities of genuine goods and make the conspicuous goods more valuable, and because of the increase of snob appeal, more people consume the genuine goods (Lai and Chang 2012).

An individual enjoys his/her own consumption but also cares for the welfare of the people that are associated with him. Altruism can be defined as individual's behavior that intended to benefit another to gain himself a higher satisfaction, even when doing so may require the payment of an implicit price. The way parents care for their children is just an altruistic example. In addition to this biological relation, altruism also can influence when making consumption decisions. Via their

consumption choice, consumers can express their concern about whether a company deals business ethically. A significant subset of consumers is the ethical consumer who feels responsible toward society and expresses his/her or her altruism by means of his/her or her purchasing behavior. The purchase of a product is associated with consumers' altruism ethic if they concern a certain ethical issue (human rights, labor conditions, animal well-being, environment, etc.) as they make consumption decisions. Ethical consumer behavior is associated with the consumptions that intend to either benefit the natural environment (e.g., environmentally friendly products, legally logged wood, animal well-being) or help people (e.g., products free from child labor, Fair trade products). Some altruistic consumers will buy the products that have specific positive qualities (e.g., green products) to show their ethical concerns (Pelsmacker et al. 2005). For example, a considerable number of consumers have shown a willingness to pay a premium for products labeled as "Fair Trade" and a favorite to retailers that are seen to be more generous to their suppliers and employees, domestically and internationally. Marketing appeal to the consumer's altruism is the most important feature of Fair Trade products. The more an individual consumer's altruism, the higher the marginal altruistic value from giving more profit to the farmers. Raising awareness of Fair trade may promote individual's altruism and result in ethical consumption. Because of Fair trade, farmers can receive greater profits, which in turn induce farmers to invest more in producing their corps. Consumers' altruism may make it happen that a firm offers a Fair trade product in a competitive environment (Reinstein and Song 2012).

Not only can an individual's own tastes and habits be the determinants of his/her decisions on consumptions, but also social activities have effects on his/her demands for some goods. Consumers make their consumption choices under social influences, when they take up the attitude of others around them without being aware of they being doing so. Derived from a cause-consequence of social influence that is not technically measurable so far, the argument is much weaker in the power of predicting consumers' choice than the one based on the factors that are supported by observables, they lead to some hypotheses and hardly to some theories. However, as the technology of data mining progresses and more resources have been invested in testing the unobservables, the theories of consumers' choices will be enriched and their predictions will be more powerful in the future.

References

Becker GS (1991) A note on restaurant pricing and other examples of social influences on price. J Polit Econ 95:1109–1116

Becker GS, Mulligan CB (1997) An endogenous determination of time preference. Q J Econ 112:729–758

Becker GS, Murphy KM (1988) A theory of rational addiction. J Polit Econ 96:675–700

Epstein GL (1987) The global stability of efficient intertemporal allocation. Econometrica LV:329–355

Grossman GM, Shapiro C (1988) Counterfeit-product trade. Am Econ Rev 78:59–75

Higgins RS, Rubin PH (1986) Counterfeit goods. J Law Econ 29:211–230

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–292

Lai FT, Chang SC (2012) Consumers' choices, infringements and market competition. Eur J Law Econ 34:77–103

Laibson D (2001) A cue theory of consumption. Q J Econ 116:81–119

Leibenstein H (1950) Bandwagon, snob, and veblen effects in the theory of consumers' demand. Q J Econ 64:183–207

List JA (2003) Does market experience eliminate market anomalies? Q J Econ 118:41–71

Lucas R, Stokey N (1984) Economic growth with many consumers. J Econ Theory XXXII:139–171

Mas-Colell A, Whingston MD, Green (1995) Microeconomic theory. Oxford University Press, New York

Pelsmacker PD, Driesen L, Rayp G (2005) Do consumers care about ethics? Willingness to pay for Fair trade coffee. J Consum Aff 39:363–383

Pindyck RS, Rubinfeld DL (2012a) Microeconomics. Pearson, Taipei

Pindyck RS, Rubinfeld DL (2012b) Difference Preferences toward Risk, 5.2 preferences Toward risk, Chapter 5, pp 159–161

Reinstein D, Song J (2012) Efficient consumer altruism and Fair trade products. J Econ Manag Strateg 21:213–241

Uzawa H (1968) Time preference, consumption function, and optimum asset holding. In: Wolfe JN (ed) Value, capital, and growth: papers in honor of Sir John Hicks. University of Edinburgh press, Edinburgh

Varian H (1992) Microeconomics. Norton, New York

Contest Theory

► Tournament Theory

Contract, Freedom of

Péter Cserne
University of Hull, Hull, UK

Abstract

Freedom of contract is a principle of law, expressing three related ideas: parties should be free to choose their contracting partners ("party freedom"), to agree freely on the terms of their agreement ("term freedom"), and where agreements have been freely made, parties should be held to their bargains ("sanctity of contract"). This entry provides an overview of the economic justifications and limitations of this principle.

Freedom of Contract: Meaning and Significance

Freedom of contract is a fundamental principle of most modern contract laws, expressing three related ideas: parties should be free to choose their contracting partners ("party or partner freedom"), to agree freely on the terms of their agreement ("term freedom") and where agreements have been freely made, parties should be held to their bargains and contracts should be enforceable by state institutions ("sanctity of contract") (Brownsword 2006, p. 50). Freedom of contract prevails to "the extent to which the law sanctions the use of contracts as a commitment device," leaving the terms of the contract agreement to the parties (Hermalin et al. 2007, p. 18).

Freedom of contract is an ideologically charged notion which attracts strongly held political views among both defenders and critics (Craswell 2000, p. 82). "Outside the legal academy, 'freedom of contract' largely serves as a slogan for laissez-faire capitalism. Even within contract theory, the term retains a particular libertarian flavor" (Dagan and Heller 2013, p. 1).

Historical research has established that the idea of a general enforceability of agreements comes from late medieval and early modern theological and philosophical debates on the moral foundations of contract law. Ancient and medieval laws did not recognize the general enforceability of consensual agreements; only certain types of agreements based on consent were enforceable. Later, both the general principle of freedom of contract and its limits have been systematically discussed in late scholastic natural law theories, thus providing moral underpinning for the rise of economic freedoms (Gordley 1991; Decock 2013).

The philosophical discussion is still ongoing as to which exchanges are morally permissible, which are problematic, and why. Yet there seems to be a reasonably broad consensus that at least a limited version of freedom of contract may be supported by both autonomy-based and welfarist theories, and, perhaps less prominently but also importantly, by aretaic (virtue-based) arguments (Cserne 2012, pp. 82–89).

Legal doctrines of most modern legal system reflect this overlapping consensus to a considerable extent. Modern Western legal systems attach a high value to freedom of contract as a basic legal principle but also set several limits to this freedom, going well beyond punctual exceptions. In these countries, "most contracts that support legitimate economic exchange are at least presumptively enforceable. Still, the limits of freedom of contract vary among Western countries and are an important element of regulatory policy" (Hermalin et al. 2007, p. 19). In fact, today's contract law regimes can be seen as long lists of exceptions to the principle of contractual freedom.

These exceptions have all been subject to analysis in mainstream law and economics scholarship: some make more (economic) sense than others. From an economic perspective, the presumption for freedom of contract is supported by its conduciveness to welfare and can be typically rebutted by identifying a bargaining failure or a market failure (Cooter and Ulen 2012, p. 341).

This, in turn, suggests either imposing mandatory terms on contracting parties or refusing the enforcement of their agreement. This functional linking of particular rules and doctrines to their incentive effects is a microlevel economic analysis of the limits of contractual freedom which provides valuable contribution to both economic and legal scholarship.

The Economic Case for Freedom of Contract

The operation of a modern market economy relies on freely negotiated enforceable contracts. Overall, in mainstream economic theory freedom of contract, sometimes under the label of consumer sovereignty (Persky 1993), has been traditionally supported by its likely benefits in terms of social welfare.

This case for freedom of contract is based on a contingent empirical generalization: "Most people look after their own interests better than anyone else would do for them" (Cooter and Ulen 2012, p. 342). In neoclassical economics, the "predilection for private ordering over collective decision-making is based on a simple (perhaps simple-minded) premise: if two parties are to be observed entering into a voluntary private exchange, the presumption must be that both feel the exchange is likely to make them better off, otherwise they would not have entered into it" (Trebilcock 1993, p. 7).

Freedom of contract and a competitive market economy seem to simultaneously promote individual autonomy and social welfare, converging toward what could be called the private ordering paradigm. According to this "convergence claim," freedom of contract is supported by a combination of autonomy-based and welfare-based arguments (Pincione 2008). Mainstream economics claims that promises should be enforced when they provide an ex ante Pareto improvement, i.e., if and only if promisor and promisee both benefit from the agreement. This is often assumed to be equivalent to the assumption that both parties wanted the agreement to be enforceable when it was made. As we shall see below, however, this convergence is not universal. In case of divergence, economists tend to give priority to social welfare considerations. "An economic case for or against freedom of contract is based on the consequent welfare implications" (Hermalin et al. 2007, p. 21).

In a perfectly competitive market, there should never be inefficient contract terms. Therefore, there is no way to improve efficiency by forbidding certain terms. This case is a useful benchmark in the sense that when one or more of the conditions of market perfection is not fulfilled, there is potential for improving efficiency by restricting freedom. In other words, circumstances when these conditions do not prevail provide an economic justification for rules and doctrines of contract regulation. The two main cases for regulating contract are then third-party effects and bargaining (contracting) failures.

Welfare economics provides theoretical justification for freedom of contract by reference to a general equilibrium economy (the First Theorem of Welfare Economics) or at least the efficiency of singular competitive markets (Hermalin et al. 2007, pp. 21–30). The Coase theorem suggests that freedom of contract is desirable even more generally. Costless contracting and the parties' rationality guarantee that they will exhaust all possibilities for mutually beneficial exchanges, thus bringing about maximum welfare.

When we relax the zero transaction costs assumption of the Coase theorem, there might be situations with positive transaction costs that justify either default or mandatory rules. Yet the transaction costs of contract regulation need to be taken into account as well. "In other words, while it is true that restrictions on private contracts can possibly enhance efficiency when the private parties incur transactions costs, one must assess that observation in light of real-life limitations on what the legal system can do and the cost at which it can do it" (Hermalin et al. 2007, pp. 27–28).

In fact, contract law (and more generally, third-party enforcement) is one among many governance mechanisms for private transactions; its use varies historically and cross-culturally. The importance of law (courts and other state

institutions) in contract enforcement depends on its merits and costs relative to other governance mechanisms (Dixit 2004). On the one hand, contract law does not seem necessary for an exchange economy to operate. As Piccione and Rubinstein (2007) showed, many equilibrium features of competitive markets (an exchange economy) can be achieved even "in the jungle," i.e., in an economy without property rights and freedom of contract where resource allocation is regulated by physical strength. On the other hand, there is ample evidence that legal enforcement of contracts is not sufficient for the welfare benefits of competitive markets to be realized. Well-functioning markets rely on social norms and informal institutions in various ways.

Economic analysis thus provides a prima facie justification for freedom of contract and also "suggests two potential grounds on which to argue against (complete) freedom of contract: (i) actors who are not party to a contract (third parties) are affected by externalities resulting from the contract; and (ii) problems in negotiating a contract prevent the parties from writing the optimal contract" (Hermalin et al. 2007, p. 30).

Constitutive, Procedural, Informational, and Substantive Limits to Freedom of Contract

Contract law is understood here as a body of legal rules that pertains to the enforcement and regulation of voluntary private agreements. Contractual freedom is not only a matter for contract law, however. It may be limited by rules outside the domain of contract law, for instance, when anti-discrimination laws constrain parties' freedom to choose their contracting partners.

The distinction between state regulation and state enforcement (negative and affirmative government sanction) needs to be noted as well: "there are many agreements that cannot be enforced in the courts but that can still be useful as commitment devices if the parties can manage to implement them privately" (Hermalin et al. 2007, p. 19).

In what follows our main focus will be on freedom of contract and its limits as they appear in rules, principles, and doctrines of contract law. The rules and doctrines of contract law have been analyzed extensively in terms of their impact on social welfare. These limits are sometimes classified into constitutive, procedural, informational, and substantive limits to freedom of contract (Cserne 2012, pp. 93–135). With respect to each, the question for economic analysis is whether and how the legal instrument in question can be illuminated, explained, justified, or criticized in light of the empirical findings and normative criteria of economics.

Contracting practices and the private ordering paradigm implicitly assume some "constitutive limits" on freedom of contract (Kennedy 1982). This term refers to those minimal conditions of individual rationality and voluntariness which are necessary for the working of even a libertarian (unregulated) contract regime. Virtually all legal systems impose threshold conditions for the making of enforceable contracts, requiring capacity, and prohibiting duress and fraud. The key idea is that these "limits" constitute our idea of what contracts are, rather than constraining freedom of contract. These constitutive limits of freedom of contract not only guarantee contracting as a domain of individual autonomy but are also instrumental for increasing social welfare (Cserne 2012, pp. 93–106). While the importance of constitutive limits is quite intuitive, incorporating them into economic models is not straightforward. "Economists frequently extol the virtues of voluntary exchange, but economics does not have a detailed account of what it means for exchange to be voluntary" (Cooter and Ulen 2008, p. 12).

Procedural limits do not constrain the parties' agreement on terms they choose (accept or bargain for), but merely require certain actions to be taken (or not taken) before contracting or during the contractual relationship. They relate to the process of agreeing on a contract and the manner of recording or authenticating the agreement. For instance, they prescribe written form, waiting period, mandatory advice, or mandatory withdrawal rights (cooling off periods).

Informational limits regulate the information flow between the parties before or during the contract. These include rules that mandate the precontractual furnishing of information and prohibit the provision of fraudulent, misleading, or irrelevant information.

Substantive limits set mandatory terms for the contracts either directly, e.g., by regulating interest rates and other terms of consumer credit contracts by statutory rules, or indirectly, e.g., by nonenforcement of terms that courts find unconscionable, unreasonable, or unfair.

Bargaining Failures and Market Failures

Cooter and Ulen's textbook treatment of regulatory doctrines of contract law (Cooter and Ulen 2012, Tables 9.3 and 9.5) can be seen as a systematic translation or linking exercise between various shortcomings of a perfectly functioning market on the one hand and the respective contract law doctrines (formation defenses and performance excuses) triggered or justified by a market failure on the other.

They focus on two kinds of shortcomings. The first include failures of the bargaining process that prevent welfare maximization (or make it unlikely) and are further classified as cases of *bounded rationality* (lack of stable and well-ordered preferences), addressed by the rules on (in)capacity, or cases of *constrained choice sets*, addressed by the doctrines of duress, necessity, or impossibility. When a contracting party faces dire (as opposed to moderate) scarcity and the other party either generates or takes advantage of this situation, the resulting contract is often not enforced. Constitutive limits to freedom of contract are always triggered in this context. Whether empirical evidence on bounded rationality justifies further limitations to freedom of contract is discussed below.

The second kind of shortcomings includes market failures that can be categorized according to three types of transaction costs which arise, respectively, from spillovers (externalities), information imperfections, and market power. These market failures are in turn addressed by various contract law doctrines as well as noncontractual regulations (Cooter and Ulen 2012, pp. 341–372).

Externalities justify the unenforceability of contracts which derogate public policy or violate a statutory duty. This category will be analyzed further below.

Symmetrical or asymmetrical *information imperfections* also generate market failures (Trebilcock 1993, Chaps. 5 and 6). These are addressed by contract doctrines such as frustration of purpose or mutual mistake (in case of symmetric imperfections) and fraud, unilateral mistake, or failure to disclose (in case of information asymmetry). Information asymmetry has been the subject of economic analysis at least since Akerlof's (1970) seminal work, and the findings generally suggest that "such distortions must imply a loss of welfare vis-à-vis the symmetric-information benchmark. [. . .] Whenever the parties negotiate imperfect contracts, the question arises whether there is scope for the legal system to improve matters, either by restricting the set of possible contracts ex ante or through appropriate court action ex post" (Hermalin et al 2007, p. 34).

The third type of market failure includes structural or situational *monopoly* or at least significant market power to restrict competition. "Competitive markets can be expected to maximize welfare in the absence of externalities. When, however, one or more entities have market power, the market can no longer be expected to yield the social welfare-maximizing allocation" (Hermalin et al. 2007, p. 39). Market power is addressed primarily by competition law, but situational monopolies also trigger contract law doctrines such as necessity and unconscionability.

Externalities, Simple and Subtle, Justifying Intervention

Externalities impose costs or benefits from a particular exchange transaction on third parties who are not involved in the transaction. Positive externalities pose incentive problems, leading to a suboptimal quantity of the good or transaction in question. Negative externalities are arguably more important with respect to contract regulation.

If such an effect can be detected, this provides reason for interfering with contractual freedom.

> The efficiency of markets and private contracting is contingent on there being no third-party externalities. For instance, the market equilibrium with a competitive, but heavily polluting, industry does not maximize welfare—the supply of the good in question is determined by the private costs incurred by the manufacturers rather than the social costs that account for both those private costs and the harm the pollution imposes on society. Because social costs are greater than private costs, more than the welfare-maximizing quantity gets sold. [. . .] More generally, in a market, bilateral contracts may generate externalities and reduce the welfare on an aggregate level. [. . .] The inefficiency of the market when externalities are present can justify restrictions on private contracts. (Hermalin et al. 2007, p. 30)

Although in principle externalities could be solved by bargaining toward a grand contract including all third parties, the number of these parties may be too big and some of them could be unknown or not yet exist. A bargaining solution would often generate insurmountable transaction costs (Hermalin et al. 2007, pp. 30–31).

Some limitations on freedom of contract can be easily and plausibly justified by externalities: "antitrust authorities may frown upon contracts that have potentially harmful effects on competition (most favored nation clauses, contracts that induce predatory or collusive behavior, etc.). Contracts between a firm and a creditor may exert externalities on other creditors, either directly through priority rules in the case of bankruptcy or indirectly through the induced change in managerial incentives. The Internal Revenue Service warily investigates employment contracts that might dissimulate real income" (Tirole 1992, p. 109).

Some other limits to freedom of contract can be seen as responding to specific forms of harmful externalities. For instance, in his *Principles of Political Economy*, John Stuart Mill referred to the statutory limitation of working hours as an example of what we would call now governmental solutions to a collective action problem: "classes of persons may need the assistance of law, to give effect to their deliberate collective opinion of their own interest, by affording to every individual a guarantee that his competitors will pursue the same course, without which he cannot safely adopt it himself" (Mill 1848, Book V chapter XI § 12). He argued that even if workers as a class would prefer to work for shorter periods, they cannot achieve it without mandatory rules limiting working hours, because each would have an individual interest working longer. While a full-fledged economic analysis of such problems is rather complex, even in a competitive market, there may be cases when such interference is Pareto improving (Basu 2007). Kaushik Basu argued that in this category of cases, overriding the principle of freedom of contract can be justified within a welfarist framework, without reference to paternalism, moralism, or even autonomy.

Similarly, in a thoughtful article, Eric Posner suggested that many protective laws of modern welfare states serve to redress imbalances created by social security and welfare laws (Posner 1995). By providing a social safety net, welfare states effectively truncate the downside of financial and other risks to citizens. This regulatory environment of a welfare state has the unintended effect of encouraging socially harmful behavior, such as irresponsible spending, risky borrowing, and overindebtedness. This suggests that many seemingly paternalistic limitations on freedom of contract may be justified by harmful externalities. When individuals take on too much risk in reliance on the welfare state, they impose external costs on society. Thus, what at first looks like a rule protecting vulnerable groups may in fact be protecting the public budget.

Virtues and Vices of Reductionist Accounts

While the above cases may be plausibly analyzed in terms of externalities, the regulatory relevance of externalities is bound with problems. The issue is both theoretical and practical: what kind of externalities should contract regulation take into account? Similar to autonomy-based theories which face difficulties delimiting relevant harms that would justify coercion, welfare-based theories have difficulties in delimiting the kinds of third-party effects that would justify welfare-enhancing intervention.

First, third-party effects are pervasive: "virtually any contract may cause some external harm, [at least by] denying other potential contracting parties the opportunity to contract with the parties to the contract in question" (Shavell 2004, p. 320). If all external effects are taken into account, the private ordering paradigm is largely at an end (Trebilcock 1993, p. 58).

To be sure, from a welfarist perspective, the mere presence of an externality is not sufficient to justify limitations. "The harm to third parties must tend to exceed the benefits of a contract to the parties themselves for it to be socially desirable not to enforce a contract" (Shavell 2004, p. 320). But as externalities are ubiquitous, regulators need additional criteria to determine what kind of externalities matter and how much weight should be attached to them. Welfare-based analyses alone do not provide tools for selecting between relevant and irrelevant externalities.

For instance, so-called moral externalities refer to the fact that some people find certain conducts morally offensive or simply disgusting. Should these effects matter for contract regulation? Mainstream economics is unlikely to provide help in this matter because without further analytical tools, economics cannot properly distinguish other-regarding preferences and tangible externalities (Hatzis 2006).

Second, even if externalities are clearly tangible or measurable in monetary terms (such as costs on dependents, the social welfare system, or the public health care system), it is contestable whether such externalities provide sufficient reason for regulatory intervention. For instance, unhealthy lifestyles or risky leisure activities may have an impact on the public budget, but it is not clear whether the freedom to purchase unhealthy food should be limited for this reason alone (Trebilcock 1993, p. 75).

More generally, one may question whether all reasons for contract regulation can be fruitfully analyzed in terms of individual and social welfare.

In its simplest versions, economic arguments classify limitations to freedom of contract according to whose welfare is increased (whose losses are prevented) by nonenforcement. Steven Shavell distinguishes two rationales for legislative or judicial overriding of contracts: the existence of harmful externalities and welfare losses to the contracting parties themselves. More interestingly, he claims that this exhausts the set of valid reasons. Other justifications for nonenforcement are merely stands-in for the previous ones. Inalienability and paternalism ultimately protect, maybe in subtle or complex ways, the welfare of either the contracting parties or third parties. Consequently these rationales are reducible to the two previous ones (Shavell 2004, p. 322).

This argument is fully compatible with the methodological and substantive assumptions of mainstream economics: welfare maximization and consumer sovereignty. Having identified negatively affected third parties, specific transaction costs, and/or informational imperfections, instances of contract regulation are considered economically justified by their social welfare benefits, i.e., to the extent that they remedy such market failures. The plausibility and success of this reductionist account, however, deserves further analysis.

Some economists argue that concerns such as commodification or inalienability cannot be easily translated into welfare terms; at most, they can be modeled as specific preferences (Hermalin et al. 2007, pp. 47–48). Paternalism is also not easily translated into economic terms (Buckley 2005; Cserne 2012, Chap. 3).

Arguably, it is both an advantage and a problem for the reductionist approach that it surpasses the conflict between welfare and autonomy inherent to certain kinds of contract regulation. The attraction of such a reductionist approach comes from its simplicity or even elegance. Not only the case for freedom of contract but all of its justified limitations can be explained in relatively narrow terms by neither relaxing the rationality assumptions nor resorting to fairness arguments.

The danger is, however, that if we consider this issue from a purely welfarist economic perspective, there is no principled limit to paternalism. Indeed, as far as their normative views are concerned, economists can be anywhere on the range between hard paternalism and hard anti-paternalism. Relying on an ad hoc mixture of welfarist and autonomy-based principles, economists are ill-equipped to handle problems of

freedom of contract which arise precisely from the conflict of these two principles. If economics acknowledges some exceptions or limits to the private ordering paradigm, as it usually does in practice, then in order to justify these exceptions, "some theory of paternalism is required, the contours of which are not readily suggested by the private ordering paradigm itself" (Trebilcock 1993, p. 21).

Pragmatic Arguments and Institutional Design

Within a reductionist economic framework, limits to allegedly welfare-maximizing interventions can be added in a contingent way, with reference to empirical facts about the functioning of the institutional mechanisms that are supposed to be used for carrying out the intervention. These contingent empirical circumstances provide pragmatic arguments for freedom of contract (Cserne 2012, pp. 31–33).

Pragmatic anti-interventionist arguments draw attention to the side effects and non-intended, often counterintentional, consequences of contract regulation. These arguments are not specific to freedom of contract but need to be taken into account in designing regulatory policies. Nor do they categorically support freedom of contract. In other words, freedom of contract may be justified *faute de mieux*, by the costs of possible interventions. More specifically, pragmatic arguments refer to (1) the overinclusiveness of rules, (2) ensuing redistributive effects, (3) the lack of information, or (4) inadequate motivations of the regulators.

The pragmatic question asked here is whether the state is more or less able to prevent undesirable contracts than other mechanisms. It is worth noting that sometimes there are very few resources needed for effective state intervention: "as long as the courts are needed to enforce contracts, the contracts will not be made, and the state does not need to police the actual making of contracts and root out the undesirable ones" (Shavell 2004, p. 322).

This leads us further into questions of institutional design. Some market failures cannot be appropriately addressed judicially, i.e., through private law constraints on freedom of contract but there may be other regulatory tools available. Ultimately, economics is likely to suggest a mix of policy instruments for contract regulation. For instance, Trebilcock argues for a "relative institutional division of labor" in which "the common law of contracts will be principally concerned with autonomy issues in evaluating claims of coercion, antitrust and regulatory law [with] issues of consumer welfare, and the social welfare system [with] issues of distributive justice" (Trebilcock 1993, p. 101).

Behavioral Economics and Freedom of Contract

At first sight, behavioral economics seems the right way to go, not only with regard to contractual freedom but policymaking more generally. By testing the assumptions of economics empirically, making economic theories descriptively more precise, one can expect to make economics more useful for policy design. Behavioral economics has identified and/or provided evidence for the functioning of various techniques of choice architecture (sticky default rules, options, menus, information provision) which take into account empirical findings on human decision making and thus can be put to socially beneficial use in contract regulation.

What else has behavioral economics to tell about freedom of contract? It is sometimes argued that psychological research provides new arguments for limiting freedom of contract. This idea seems misconceived (Cserne 2012, pp. 43–54, 137–139). There are distinct economic arguments for limiting freedom of contract in certain circumstances, and empirical research is indispensible for identifying whether and to what extent these circumstances prevail. Also, empirical findings give more detail and in this respect further support to the existing argument about the imperfection of the correlation between individual choice and welfare maximization. The increased attention to these findings is expected to lead to more precise and better-founded knowledge about the circumstances when contracting parties make

suboptimal choices. Yet, the evidence on various cognitive biases does not, in itself, call for limiting freedom of contract any more or less than common sense observations about human frailties. Indeed, it is an open question whether behavioral findings justify more or less paternalistic regulation than we currently can observe. (Note that the relevant comparison is this, not the one between hard paternalism and a hypothetical libertarian "regulatory" regime.) Some empirical research draws attention to psychological advantages of giving people the freedom to choose (Feldman 2002). More importantly, the normative standards for contract regulation need to come from elsewhere than empirical research. In this regard, behavioral economics does not fare any better or worse than mainstream economics.

Cross-References

- ▶ Liberty
- ▶ Limits of Contracts
- ▶ Market Failure: Analysis
- ▶ Market Failure: History
- ▶ Public Goods
- ▶ Rationality

References

Akerlof GA (1970) The market for lemons: quality uncertainty and the market mechanism. Quarterly J of Economics 84:488–500

Basu K (2007) Coercion, contract and the limits of the market. Soc Choice Welfare 29:559–579

Brownsword R (2006) Freedom of contract. In: Brownsword R (ed) Contract law. Themes for the twenty-first century, 2nd edn. Oxford University Press, Oxford, pp 46–70

Buckley FH (2005) Just exchange: a theory of contract. Routledge, London

Cooter R, Ulen T (2008) Law and economics, 5th edn. Pearson Education, Boston

Cooter R, Ulen T (2012) Law and economics, 6th edn. Pearson Education, Boston

Craswell R (2000) Freedom of contract. In: Posner E - (ed) Chicago lectures in law and economics. Foundation Press, New York, pp 81–103

Cserne P (2012) Freedom of contract and paternalism. Prospects and limits of an economic approach. Palgrave, New York

Dagan H, Heller MA (2013) Freedom of contracts. Columbia law and economics working paper no. 458. Available at SSRN. http://ssrn.com/abstract=2325254

Decock W (2013) Theologians and contract law. The moral transformation of the Ius Commune (ca. 1500–1650). Martinus Nijhoff, Leiden

Dixit A (2004) Lawlessness and economics. Alternative modes of governance. Princeton University Press, Princeton

Feldman Y (2002) Control or security: a therapeutic approach to the freedom of contract. Touro L Rev 18:503–562

Gordley J (1991) The philosophical origins of modern contract doctrine. Clarendon, Oxford

Hatzis AN (2006) The negative externalities of immorality: the case of same-sex marriage. Skepsis 17:52–65

Hermalin BE, Katz AW, Craswell R (2007) Contract Law. In: Polinsky AM, Shavell S (eds) The handbook of law & economics, vol 1. Elsevier, Amsterdam, pp 3–136

Kennedy D (1982) Distributive and paternalist motives in contract and tort law, with special reference to compulsory terms and unequal bargaining power. Maryland Law Rev 41:563–658

Mill JS (1848) Principles of political economy with some of their applications to social philosophy. http://www.econlib.org/library/Mill/mlP.html

Persky J (1993) Consumer sovereignty. J Econom Perspect 7:183–191

Piccione M, Rubinstein A (2007) Equilibrium in the jungle. Econom J 117:883–896

Pincione G (2008) Welfare, autonomy, and contractual freedom. In: White MD (ed) Theoretical foundations of law and economics. Cambridge University Press, Cambridge, pp 214–233

Posner EA (1995) Contract law in the welfare state: a defense of the unconscionability doctrine, usury laws, and related limitations on freedom of contract. J Legal Stud 24:283–319

Shavell S (2004) The foundations of economic analysis of law. Harvard University Press, Cambridge

Tirole J (1992) Comments. In: Werin L, Wijkander H (eds) Contract economics. Basil Blackwell, Oxford, pp 109–113

Trebilcock MJ (1993) The limits of freedom of contract. Harvard University Press, Cambridge

Further Reading

Atiyah PS (1979) The rise and fall of freedom of contract. Clarendon, Oxford

Ben-Shahar O (ed) (2004) Symposium on freedom from contract. Wisconsin Law Rev 2004:261–836

Buckley FH (ed) (1999) The fall and rise of freedom of contract. Duke University Press, Durham

Craswell R (2001) Two economic theories of enforcing promises. In: Benson P (ed) The theory of contract law: new essays. Cambridge University Press, Cambridge, pp 19–44

Kerber W, Vanberg V (2001) Constitutional aspects of party autonomy and its limits: the perspective of

constitutional economics. In: Grundmann S, Kerber W, Weatherhill S (eds) Party autonomy and the role of information in the internal market. De Gruyter, Berlin, pp 49–79
Kronman AT, Posner RA (1979) The economics of contract law. Little & Brown, Boston
Schwartz A, Scott RE (2003) Contract theory and the limit of contract law. Yale Law J 113:541–619

Contracts of Adhesion

Elena D'Agostino
Department of Economics, University of Messina, Messina, Italy

Abstract

In the globalized mass production economy with a large number of individual consumers, transactions very often take place between parties who are not physically present, such that communication between them turns out impossible or, at least, highly expensive. For that reason contracts are usually proposed by sellers to consumers on a take-it-or-leave-it basis without negotiation, referred to as contracts of adhesion. Consumers usually do not read the whole contract, and sellers can include inefficient one-sided clauses in fine print. This work reviews the main legal and economic literature on this topic presenting the traditional reasons to justify regulation in favor of consumers and highlighting its risks.

Introduction

According to a well-known definition, a contract is "a meeting of minds" between two or more parties who bargain on its terms and conditions. Moving from the legal definition, contracts are also viewed as a fundamental and irreplaceable tool in economics to realize an efficient allocation of limited resources among agents.

It is unanimously recognized that what should characterize an effective contract is the balance between parties' power in order to avoid that any of them may exploit some bargaining power against the other. On the contrary, a contract should be the natural conclusion of a bargaining process in which parties do not fight each other but have to reach a compromise between their opposite interests after a (more or less long and, Coasian speaking, expensive) discussion about what terms to be bound to.

If it is true that contracts are enforceable when all parties knowingly consent, nevertheless a knowledgeable consent is sufficient but not necessary to make the contract enforceable. Indeed in the globalized mass production economy with a large number of individual consumers, transactions very often take place between parties who are not physically present, such that communication between them turns out impossible or, at least, highly expensive. As a consequence, it is surprising that most of the contracts we sign (or simply accept in words) every day do not come out from a bargaining process, but every term is proposed by one party to the other, and the latter simply limits her will to adhere to the preprinted content: for this reason lawyers refer to this category of contracts as *contracts of adhesion*. Furthermore, when the same content is reproduced in every contract for the same good and proposed to any potential customer on a take-it-or-leave-it basis, the contract is not simply adhesive but also standard.

Standard contracts of adhesion characterize several markets (think of transportation, bank contracts, insurance policies, and the huge and endless list of online contracts) and are not necessarily bad contracts but rather make transactions quicker in so far as they are able to economize on some costs, like writing down the contract, finding legal references for each clause, and so on. The downside, however, is that the drafter party could exploit his bargaining power to insert one-sided clauses, that is, clauses whose content turns out very onerous for the other party and very generous for himself. To put these clauses away from the counterparty's eyes, the drafter usually relegates them at the bottom of the contract, in some annexes or footnotes, and writes them in fine print using very unfriendly legal terms to make the content

intentionally unavailable or at least obscure for nonprofessional readers.

It turns out that whether the nondrafter party reads such terms is not an open question, especially when she is a final consumer not professionally involved into the transaction: in respect to consumers, the answer to the question is quite simple, and consumers usually do not read standard contracts and, in particular, do not read fine print, so their signatures do not necessarily imply that their consent has been knowledgeable (see Bakos et al. 2014 on infrequent reading).

The reason why consumers do not read is not necessarily found in their incapacity to realize the presence of these clauses and the risk involved into signing without reading, but it could be the rational decision of not investing time and resources into an activity that may turn out useless either because it is too costly or because the consumer needs the good and knows that, even if she does not totally agree with every clause, they are unalterable and the only alternative is to reject the whole contract.

Suppose, per contra, the consumer decides to read every term. Given the obscure language used to write some or all these clauses, we have already noted that it is very unlikely that she will be able to understand their content, especially if she is not an expert in legal terms. For this reason the literature stresses on the fact that reading implies a cost on the side of consumers. It does not mean that consumers sign without having any idea of contract terms; rather it is likely that they limit their attention to some clauses only (above all those fixing price), for this reason defined as salient, and skip some other (like those regulating insurance, restoration in case of damages, place of jurisdiction, etc.), accordingly defined as non-salient.

Contracts of Adhesion in the Law Literature

The law literature contains at least two different approaches, both aiming to justify regulation of complex contracts with fine print to protect potentially unaware buyers.

The first approach is based on market structure. In this sense, Kessler (1943) argues that monopolists exploit their market power by offering contracts containing onerous terms with buyers not able to understand their content and/or to renegotiate some or all terms. While Kessler does not discuss regulation, his argument suggests that courts should be more prone to strike down standard-form contracts drafted by a monopolist. More recent versions of this approach include Kornhauser (1976), who claims that oligopolists would agree to draft onerous terms to facilitate price-fixing, and Shapiro (1995), who argues that competition would protect buyers from exploitation. Some courts have used market structure as a criterion for treating terms as procedurally unconscionable (and therefore unenforceable), notably in Henningsen v Bloomfield Motors (NJ 1960) and more recently in Pack v. Damon Corp (E.D. Mich 2004) and Flores v. Transamerica HomeFirst Inc. (Cal. Ct. App. 2001). This view corresponds to the traditional economic thought based on the supposition that a free competitive market provides efficient clauses in equilibrium: an argument that collapses if consumers have to pay a cost to read and to understand contract terms as efficiency in a competitive market requires that they are fully informed and rational.

Kessler's argument has been discredited on both empirical and theoretical grounds. Theoretically speaking, competitive firms also offer non-negotiable, complex contracts, whose terms are usually not less onerous. Empirically speaking, there are studies focusing on specific markets which demonstrate that the severity of terms included in standard-form contracts does not depend on market structure (see Marotta-Wurgler 2008 for the market for software licenses, and Priest 1981 specifically for warranty terms included in standard-form contracts). Moreover, according to a conventional argument, monopolists are better served by raising price than by including onerous terms (for a detailed analysis, see Rakoff 1983).

On an alternative account, regulation is considered a necessary step when contracts are preprinted by one of the two parties because the other party, usually consumers, should be protected as

they lack the expertise to understand and/or could be too naive to regard unread terms skeptically. A recent behavioral literature treats the infrequency of reading as indicative of consumer naivety and, in contrast to the Kessler tradition, argues that competitive sellers lack an incentive to educate such buyers: cf. Gabaix and Laibson (2006) and Gilo and Porat (2011). Regulations may therefore benefit such naive buyers, irrespective of market structure (see ▶ Naïve Consumers: Contract Economics).

Contracts of Adhesion in the Economic Literature

Various papers incorporate a cost of reading terms that are drafted by one of the parties to the contract. Katz (1990) analyzes a monopoly in which the only seller can choose non-price terms from an interval that depends on the legal regime under consideration. A monopolist chooses price that consumers observe at no cost and also whether to disclose its chosen non-price terms at some cost. If the seller does not disclose, then each buyer can either accept or read, also at some cost. In the unique equilibrium outcome, the monopolist discloses whenever this is cheap enough; otherwise, she offers the most onerous terms that the regime allows (conditional on not disclosing): the monopolist optimally sets terms such that a reading buyer never strictly prefers to accept. Katz considers a family of regulations which includes mandating favorable terms and shows that mandating the same terms as a disclosing seller would choose induces an efficient outcome. This regulation solves a commitment problem in Katz because a monopolist who did not disclose would offer the most onerous terms that the regime allows.

D'Agostino and Seidmann (2016) analyze a monopoly in which the only seller offers a menu of either complex or simple nonnegotiable contracts: the latter contains a (transparent) price and default non-price terms alone, while the former also contains a shrouded non-price term which is either favorable, default, or onerous (for all buyers). Consumers incur a (sunk) cost if they read the shrouded terms in a complex contract; and trade on favorable terms is socially efficient. If some buyers are sophisticated, then in equilibrium trade must be inefficient because of a commitment problem which is intrinsic to fine print. On the one hand, sellers cannot credibly promise that the terms in unread complex contracts are favorable and can only be disciplined to (sometimes) draft such terms if buyers read. Without such discipline, the monopolist would offer onerous contracts, which (sophisticated) consumers would then reject. On the other hand, sophisticated consumers will only incur the costs of reading if the seller randomizes over terms, and the utility difference between the best and the worst terms justifies the cost. Equilibrium play is always inefficient because trade sometimes occurs in contracts that do not include favorable terms, and sophisticated consumers then engage in socially wasteful reading.

The authors consider the effects of two sorts of regulation of non-price terms. They first compare the effects of a regulation that mandates favorable terms (that also correspond by assumption to efficient terms). Given the model they present, such regulation resolves the commitment problem, but the identification of the beneficiaries differs from Kessler's. Precisely the authors prove that resolution of the commitment problem helps the stronger side of the market in a monopoly. Thus, a regulated monopolist gains because she can now extract consumers' maximal surplus, while her sophisticated consumers are unaffected. In sum, the distributional effects of this regulation depend on market structure but in the opposite direction to Kessler's suggestion. The authors also consider the effects of a regulation that prohibits onerous terms alone. This mitigates, but does not eliminate the commitment problem; and, as a result, regulated trade is also inefficient. Indeed, such regulation may lower welfare by reducing the utility difference between the best and the worst terms in complex contracts, which drives any complex contracts out of the market by deterring consumers from reading, and thereby disciplining the seller.

Turning to a competitive market, in Che and Choi (2009) competitive sellers can decide

whether to disclose (at some cost) or not to disclose their terms that could be either favorable or unfavorable. Consumers hold heterogeneous preferences over non-price terms and can either accept or read (at some cost) the contract, possibly resampling another seller if they reject on a first instance. If sellers do not disclose, whenever consumers read in equilibrium, they are also indifferent between accepting and reading, and use price as a signal for the quality of contract terms. If sellers disclose, then consumers can immediately have access to contract terms without paying any cost. Che and Choi (2009) compare play in legal regimes with a duty to read and with a duty to speak: the former regime corresponds to an unregulated market, and sellers separate into those offering favorable terms with high probability and not disclosing and those who are sure to offer onerous terms and disclose; sellers must disclose in the latter regime but can still offer onerous contracts. Che and Choi find that the two regimes cannot be welfare ranked. However, the duty to read regime welfare dominates when, ceteris paribus, reading is cheap enough, whereas buyers are better off in the duty to speak regime if enough of them care about non-price terms. The latter regime does not result in efficient trade because sellers charge a single price to consumers with heterogeneous valuations and because of the cost of speaking.

Regulations

As pointed out in previous sections, a large consensus has grown up in the last decades among lawyers and economists that consumers must be protected against evidently unfair, non-negotiated terms. The key question in examining alternative legal systems is how protective these measures should be. Comparing different legal systems, such as the US and the EU systems, it turns out that the former opted for a regulation based on clause disclosure, whereas the latter preferred a regulation of contract terms.

What clearly emerges from the US system is, however, the lack of an organic regulation of contracts of adhesion and rather a proliferation of acts or pieces of laws applying to specific markets and/or to specific transactions. Examples can be found in the Truth in Lending Act of 1968 for the credit market, in the Magnuson-Moss Warranty Act of 1975 applying to the market for warranties, in the Nutritional Labeling and Education Act of 1994 about the food market, and, more recently, in the Principles of the Law of Software Contracts of 2009 approved by the American Law Institute. All these laws have limited application to specific markets, but looking at their rules, there is a common feature: all of them emphasize the importance of consumer's awareness of terms and conditions, and regulation mainly consists of mandating disclosure of clauses written in fine print but leaving the drafter free to decide the content of these clauses.

This approach aims to protect the main principle in contract law, well known as freedom of contract. However, as argued by Korobkin (2003), such an intervention will sort out a positive effect if consumers are not sophisticated, that is, they are not able to understand the risk involved in signing the contract without reading every clause (see again entry on ▶ Naïve Consumers: Contract Economics). There is empirical evidence against this argument: Marotta-Wurgler (2012) shows that making it available for the consumer to access terms and conditions in online contracts for software licenses (as measured by the number of clicks required to access the corresponding window) has a negligible impact on the rate at which consumers actually read them. On the same line, Ben-Shahar and Schneider (2011) argue that to be effective, disclosure should be brief, easy, and simple, but how brief, easy, and simple it should be is still an open question.

Moving to the EU system, the 93/13/EEC Directive is the main piece of law regulating the usage of contracts of adhesion with standard clauses. On a first look, we can find some rules following the same spirit of US regulation to the extent that the aim seems to disclose at least the existence, if not the content, of some risky clauses: e.g., it is required that the consumer declares to have read terms and conditions by putting an additional signature (or by ticking a box if the transaction is online). On a second and

deeper reading, however, the approach of the EU Directive looks like pretty different from the US system. First of all, the Directive has a general application to every contract of adhesion and not just to specific markets. Secondly its main concern is evidently to look directly at the content of fine print in order to avoid their use when it turns out so one-sided to become vexatious. The Directive also contains a black list of clauses that are presumed to be vexatious and, for this reason, never enforceable even if included in the contract.

Despite these differences between the two systems, it must be reminded that courts in both systems (and also national courts for European countries) have played an important role in identifying high-risk cases and have also used the market structure criterion in order to regulate those markets where sellers exploit some market power.

Cross-References

▶ Naïve Consumers: Contract Economics

References

Bakos Y, Marotta-Wurgler F, Trossen D (2014) Does anyone read the fine print? J Leg Stud 43:1–36

Ben-Shahar O, Schneider CE (2011) The failure of mandated disclosure. Univ Pennsylvania Law Rev 159:647–749

Che Y-K, Choi A (2009) Shrink-wraps: who should bear the cost of communicating mass-market contract terms? mimeo

D'Agostino E (2005) Contracts of adhesion between Law and Economics. Rethinking the unconscionability doctrine. Springer

D'Agostino E, Seidmann DJ (2016) Protecting buyers from fine print. Eur Econ Rev 89:42–54

Gabaix X, Laibson D (2006) Shrouded attributes, consumer myopia and information suppression in competitive markets. Q J Econ 121:505–540

Gilo D, Porat A (2011) Viewing unconscionability through a market lens. William Mary Law Rev 52:133–195

Katz A (1990) Your terms or mine? The duty to read the fine print in contracts. RAND J Econ 21:518–537

Kessler F (1943) Contracts of adhesion – some thoughts about freedom of contract. Columbia Law Rev 43:629–642

Kornhauser L (1976) Unconscionability in standard forms. Calif Law Rev 64:1151–1183

Korobkin R (2003) Bounded rationality, standard form contracts, and unconscionability. Univ Chicago Law Rev 70:1203–1295

Marotta-Wurgler F (2008) Competition and the quality of standard form contracts. J Empir Leg Stud 5:447–475

Marotta-Wurgler F (2012) Does contract disclosure matter? J Inst Theor Econ 168:94–119

Priest G (1981) A theory of the consumer product warranty. Yale Law J 90:1297–1352

Rakoff T (1983) Contracts of adhesion: an essay in reconstruction. Harv Law Rev 96:1173–1284

Shapiro C (1995) Aftermarkets and consumer welfare: making sense of Kodak. Antitrust Law J 63:483–511

Contracts, Forward

Haksoo Ko
School of Law, Seoul National University, Seoul, Republic of Korea

Abstract

A forward contract is an agreement to buy or sell an asset at a specified future time at a pre-specified price. While financial underpinnings of forward contracts are well-known, law and economics research in forward contracts is underdeveloped.

Definition

A forward contract is an agreement to buy or sell an asset at a specified future time at a pre-specified price. A forward contract can be contrasted to a spot contract, which is an agreement to buy or sell an asset almost immediately. A forward contract is similar to a future contract. A main difference is that, whereas future contracts are typically standardized and trade on exchanges, forward contracts are nonstandardized and trade on the over-the-counter market.

Contract, Forward

A forward contract is an agreement to buy or sell an asset at a specified future time for

a pre-specified price (Hull 2014; Bodie et al. 2010). A party to a forward contract assumes a "long position," and the other party assumes a "short position." The party with a long position agrees to buy the underlying asset on the specified future date for a pre-specified price, while the party with a short position agrees to sell the asset on the same date for the same price.

A popular type of forward contracts is for foreign exchanges. As an illustration, imagine a European company which must purchase materials from Chinese suppliers and pay in Chinese Yuan on a regular basis. The European company would then be subject to cost uncertainty due to the unpredictable nature of the future exchange rate between Euro and Yuan. This company can reduce the uncertainty by entering into a forward contract with a bank, fixing the exchange rate at maturity in advance, regardless of the then prevailing exchange rate. Forward contracts can be entered into for all sorts of assets, whether financial or physical.

Parties enter into a forward contract for a variety of reasons. A common reason would be parties' different expectations about the future, e.g., as to whether a foreign exchange rate would rise or fall. In such a case, a forward contract would simply reflect the parties' different views or different evaluations on future circumstances in the foreign exchange market. Since the value of a forward contract at maturity will change depending on the spot price which will then be prevailing, the parties' agreement in this context can be viewed as their bet on the spot price in the future. Another reason why parties enter into a forward contract would be to hedge against fluctuations of the value of the asset that is subject to the contract. For a party, entering into a forward contract would be like buying insurance, while it would be like selling insurance for the other party. That way, the parties could reallocate and share risks.

A forward contract would show a distributional effect, largely reflecting the difference between what the parties expected at the time of entering into the contract and the actual outcome at maturity. A forward contract could also have an effect on allocation if, e.g., a party is better positioned to pool a similar type of risks together. Thus, for instance, between a bank and a manufacturing company exposed to risks arising from foreign exchange rate fluctuations, it would typically be the bank which provides the function of pooling risks.

The structure of a forward contract is often simple and straightforward. As such, so long as there are no disputes as to the validity of the forward contract and also as to the applicable contract terms, contract obligations would become clear at maturity. Thus, if there are disputes between the parties, it may well be regarding the inherent validity of the contract, rather than regarding the interpretation of specific contract terms.

In that context, a forward contract could prove to be problematic, if a party has sophisticated expert knowledge about the relevant market, while the other party lacks such knowledge. In such a case, depending on the market situation at maturity, the party claiming the lack of knowledge may refuse to carry out contract obligations and may challenge the validity of the contract. In challenging the validity of the contract, this party may claim that the forward contract is a fundamentally unfair contract and allege violation of the general legal principle requiring good faith when entering into a contract. Fraud, mistake, failure to explain, and various other legal doctrines could also be cited in challenging the validity of the forward contract.

However, without the detailed factual information surrounding the parties' dealings at the time of entering into the forward contract, it would be difficult to assess whether the party's claim challenging the validity of the forward contract is itself made bona fides or whether the party is engaging in ex post opportunistic behavior.

In assessing the parties' judgment when entering into a forward contract, a behavioral law and economics perspective could be helpful. Behavioral law and economics could shed light on the parties' motivations and behavior at the time of entering into contract and could help in learning if psychological biases and limitations such as investor myopia, optimistic bias, and herd mentality had impact on one or both parties (Ko and

Moon 2012). In particular, if a contracting party lacks sophistication, enhancing understanding as to how individual contract provisions were inserted into a specific forward contract would help in assessing whether the forward contract could be viewed as a result of arm's-length dealings. In entering into a forward contract, the parties often use a standard form contract, and as such, the law and economics literature on standard form contracts could be helpful as well.

Overall, law and economics research in forward contracts is underdeveloped. As we learn more about what precisely transpires when the parties enter into a forward contract, academic research in this area will become richer and more interesting.

References

Bodie Z, Kane A, Marcus A (2010) Investments, 9th edn. McGraw-Hill, New York

Hull JC (2014) Options, futures, and other derivatives, 9th edn. Prentice Hall, New Jersey

Ko H, Moon W (2012) Contracting foreign exchange rate risks: a behavioral law and economics perspective on KIKO forward contracts. Eur J Law Econ 34:391–412

Cooperative Game and the Law

Samuel Ferey
CNRS, BETA, University of Lorraine, Nancy, France

Definition

While noncooperative game theory applied to the law is now a subfield of law and economics literature, cooperative game theory has a more strange history. Many of the founding fathers of the cooperative game theory (Shapley, Shubik, Owen, Aumann) were interested in legal examples to illustrate their games; however, law and economics literature has not systematically investigated the meaning of cooperative game theory for the law and is still mostly noncooperative oriented. The aim of the entry is to draw a general picture of what cooperative game theory may add to the law and economics literature. We focus on the positive and normative aspects of cooperative game theory, and we provide illustrative examples in different fields (private law, public law, regulation, theory of the law).

Cooperative Game Theory and the Law: A Missed *Rendezvous*?

In their famous book *Game Theory and the Law* published in 1994, Baird, Gertner, and Picker said nothing about cooperative games and the law (Baird et al. 1994). The authors mainly focused on noncooperative games and gave to the Nash equilibrium the most preeminent role to understand strategies of legal players (plaintiffs, defendants, and judges). Most of the subfields of law and economics have been deeply changed by the use of game theory in place of the Chicago-style price theory. Compared with noncooperative game theory, cooperative game theory is still underestimated. From a historical perspective, here lies a paradox. While most of the classical games studied by cooperative branch of game theory have obvious consequences for the law (the ownership game by Shapley and Shubik 1967; the bankruptcy game by Aumann and Maschler 1985; the airport games by Littlechild and Thomson 1977), no structured paradigm on law/economics/cooperative games has emerged compared to the noncooperative game approach. The first reason lies in the fact that the legal examples modeled with cooperative game theory were not explicitly oriented in a law and economics perspective. The second reason is that, at the very beginning of the law and economics movement, leaders of the field, like Posner or Calabresi, were almost exclusively interested in the efficiency of the common law (maximization of wealth) which is not an issue addressed by cooperative game theorists.

Much has changed in recent times. A lot of new and original models focus on unprecedented applications that can be envisaged through cooperative game theory as applied to private law, public law, regulation, and legal theory. These

legal-oriented models may overlap applications in public economics or industrial organization (imperfect competition, public goods, matching, and networks). We have chosen to focus on some of the most suggestive applications of cooperative game theory for the law.

As we would like to avoid non-useful technical complexities, we deal with the most basic and simple games to show how they renew our ideas on what economics may add to the law. More importantly, we insist on the twofold features of cooperative game theory which can be considered from a positive point of view (how people cooperate and share the surplus created by cooperation) and from a normative view (how a judge or an arbitrator should settle a case when several people are in conflict about how to share a joint surplus).

First, definition and notation are introduced in order to better understand what cooperative game theory is. Second, we deal with more contemporary works using cooperative game to highlight regulation and public law issues. Third, we show that cooperative game approach is useful to renew private law and mainly torts. Last, we give some intuitions at a more abstract level to see how the legal theory could be influenced by the cooperative game approach.

Definition, Notation, and Meaning

A transferable utility game (TU game) is a couple (N,v) with N the set of players and v the characteristic function. The two basic blocks of cooperative game theory are the coalitions and the characteristic function. The Grand coalition $\{1, 2, \ldots n\}$ is the coalition of all the players. The singletons $\{i\}$ are coalitions restricted to single players. There are also all the subsets of players between singletons and N. With n players, there are $2^n - 1$ coalitions. The worth of the Grand coalition is $v(N)$ and is equal to the worth to be shared among players. The worth $v(i)$ is what player i could get if he decided to behave on his own. More generally, the worth $v(S)$, with S a coalition of players, is what the coalition S is able to get for its members. The characteristic function associates to each coalition its worth but says nothing about how this worth is then shared among the members of S. Mostly, a hypothesis of transferability holds. Transferability is a serious assumption and states that utility is measurable in terms of money: side payments are possible among the players.

Once coalitions and characteristic function are defined, then the most important challenge for cooperative game theory is to solve the game that is to say to find a vector of payment $(x_1, x_2, \ldots x_n)$ for the players. At first sight, infinity of vectors could work. But, two rationality requirements will be added. First is collective rationality condition: the sum of the payments should be necessarily equal to the worth of the Grand coalition (no more, no less). The vector of payment should be feasible (it is impossible to give players more than what is created by the Grand coalition) and should ensure that resources are not wasted (it would be irrational to give them less than the surplus jointly created). Second is the individual rationality condition: the payment of a player i should be more important that its worth $v(i)$: without this condition, a single player would have no incentive to cooperate with others. The vectors which satisfy these two conditions are called imputations.

Then, a first set of solutions is the core. The core of a game $C(N,v)$ is defined as follows:

$$C(N,v) = \{x \in \mathbb{R}^n \mid x(N) = v(N) \text{ and } x(S) \geq V(S) \text{ for all } S \subset N\}$$

Behind the core is the idea of stability. In case of non-empty core, no individual nor any coalition has an interest to leave the Grand coalition for their own. On the contrary, an empty core means no guarantee on the stability of cooperation among players: some of them have an incentive to leave the Grand coalition and get more, out from the Grand coalition. Cooperative game theorists have then studied the conditions under which the core is non-empty (e.g., when games are convex – that is to say $v(S) + v(T) \leq v(S \cup T) + v(S \cap T)$ for all S and T – the core is non-empty). Still, the set of the core may be large enough.

It is possible to define other concept solutions with an axiomatic perspective. A trade-off arises between "not enough" and "too much" axioms.

If few axioms are required, it will be easy to find allocations which solve the game, but the set of solutions is likely to be too large. If too much axioms are required, the set of solution is likely to be empty. Each axiom is debatable on rational and normative grounds.

A particularly interesting rule to solve a game is the Shapley value. The Shapley value may be explained in two alternative ways. First, the Shapley value depends on the marginal contribution of players to the coalitions. For a player i, its marginal contribution to a coalition S is the difference between the worth of the coalition $v(S)$ and the worth of the coalition $v(S\backslash i)$. The Shapley value allocates to player i his average marginal contribution. Second, the Shapley value – defined on any cooperative game – follows four axioms which characterize uniquely this rule. The first is efficiency (the worth of the Grand coalition should be shared); the second is the null player axiom (players who contribute zero to any coalition get nothing back); the third is the additivity axiom (for a game that is the sum of two games, the Shapley value for the former is the sum of the Shapley values for the latters); and the fourth is symmetry (two substitutable players should receive the same payoff).

Beyond the core and the Shapley value, other sharing rules are discussed in the literature (e.g., the nucleolus). More recently, new paths in cooperative game theory have been developed: games with a priori unions (Owen 1977) or graph-restricted games (Myerson 1977). In these games, a structure of cooperation is defined through a network of preexisting links, and the value is calculated on this structure.

Cooperative Games, Antitrust, and Regulation: From Stability to Fairness

A contemporary perspective on cooperative game theory and the law would consider that public regulation is one of the main fields where cooperative game theory is useful. In part I, we have shown that cooperative game theory may be considered from a twofold perspective: first, it models "situations in which the players may conclude binding agreements that impose a particular action or a series of actions on each player" (Maschler et al. 2013); second, axiomatization of solution concepts indicates how a judge or an arbitrator should allocate the value among the players. The first perspective could be said positive and the second normative. As soon as public agencies aim at regulating the behaviors of individual or firms implied in a common activity, cooperative game theory has something to say about the following: (1) Is cooperation among players stable? (2) How will be (or should be) the surplus due to cooperation shared?

A first subfield is antitrust. Collusion and anticompetitive practices may be considered as cooperation among players: colluders coordinate their actions in order to get monopoly profits. Some of antitrust scholars assert that cooperative game theory is irrelevant insofar as collusion being illegal, there is no way to conclude binding agreements; on the contrary others consider that the non-emptiness of the core is useful to better understand the stability of cartels and the efficiency of anticompetitive practices (Telser 1985; criticized in Wiley 1987). Analyzing the famous *Addyston Pipe* case, one of the most famous cases in antitrust according to Judge Bork, Bittlingmayer and Telser argue that cartelization may be useful for the firms to share fixed costs (Bittlingmayer 1982). This cooperation avoids the inefficiencies in sectors where marginal cost is under the average cost. This idea is followed by contemporary literature on oligopoly games. Both Cournot and Bertrand oligopolies are concerned. The main issue addressed is the stability of collusion. Some of anticompetitive behaviors (as prohibitive restrictive agreements or concerted practices) imply that players coordinate their strategies to get their best outcome and may organize side payments. In some models, a hypothesis of sharing the best technology is done, but it is not a necessary condition (Lardon 2017). The stability of the agreement among members of the oligopoly is one of the key elements of this literature. The core is consequently the most studied solution: if the core is non-empty, stability of the cartel is expected; on the contrary, the emptiness of the core implies that the cartel is unstable insofar as some players or groups of players have an incentive to leave the Grand coalition. Zhao applies this

reasoning to the sugar cartel in the USA at the end of the nineteenth century. For this author, the increasing of the cartel organizational costs due to the Sherman Act made the core empty and lead to instability (Zhao 2014). What is interesting in this literature of oligopoly games is to work out different scenario implying different blocking rules. Indeed, when a player or a group of players decide to leave the Grand coalition, the interaction between them and the remaining players becomes strategic, and several strategies are conceivable. For example, they may behave to maximize their own utility or to minimize the utility of the other players. The conclusions drawn are particularly interesting for antitrust authorities regarding the best way to enforce antitrust law and to enhance instability of collusion.

A second subfield is concerned with public regulation, public utilities, and facilities at large. In many contexts covered by public regulation, agents (firms, individuals, or groups like municipalities) cooperate and are looking for the best way to share their costs: fisheries conferences, commons, pools, water or telecom networks, facilities that will be jointly used, and water from a river at a national or international level are some of the numerous examples of such situations. Take the example of the airport fees analyzed by Littlechild and Thomson (1977) which deals with how to share the cost of a common facility (a runaway). Assume three aircraft companies which cooperate to build a new runaway. Due to the size of the planes they own, the first company needs a small airstrip, the second a medium one, and the third a large one. The cost of each airstrip is $c1 < c2 < c3$. If firms do not cooperate, they bear a total cost of $c1 + c2 + c3$, while cooperation leads to a total cost equals to $c3$ (we suppose that there is no congestion and the largest airstrip is enough to land all the aircrafts whatever their size is). The core is non-empty but large enough and let unsolved the precise cost paid by each firms. A regulator could use more specific sharing rule. A natural solution would say that the total cost $c3$ should be divided as follows: the smallest part of the airstrip is used by all the companies: $c1$ should be equally divided among the three companies. The additional cost $(c2-c1)$ should be paid only by the companies 2 and 3. The additional cost $(c3-c2)$ should be paid by the third firm (the only firm which needs a large airstrip). This intuitive solution is the Shapley value of the airport game. As such, the Shapley value is a fair compromise that a regulator could implement to sharing the costs among firms.

Sometimes, legislators and regulators play a major role in cooperation. This is the case when regulated firms are required by the law to cooperate. The REACH legislation in the European Union (Registration, Evaluation, Authorization and Restriction of Chemicals) is very illustrative. The European Union has decided in 2006 to make a systematic evaluation of the toxicity of chemicals. The number of chemicals is so high that it is impossible for public authorities to evaluate by themselves their danger. At the same time, firms privately own information regarding the toxicity of chemicals (scientific reports, experiences, etc.). REACH legislation creates SIEF which are forum to organize the exchange of information among firms. The difficulty lies in the side payments: some firms may provide very valuable data, some add information already get by others, some have no information at all, etc. Based on the costs of replication of data, Dehez and Tellone (2013) provide a cooperative game model and advocate the Shapley value as a fair compromise to calculate payments/rewards of each participant to a SIEF. Beal and Deschamps follow this path and discuss different rules in order to compare their axiomatic properties (Beal and Deschamps 2016).

Last, and more recently, a third important subfield using cooperative game theory has grown: the implementation of public matching algorithms. Regulators may face complex issues of matching the two sides of the market. In a famous paper, Roth studied the National Resident Matching Program in the USA that aims at matching hospitals with resident doctors (Roth 1984). The NRMP is a centralized "clearinghouse" introduced in the 1950s to allocate resident doctors to hospitals. Roth discovered that the algorithm used in the NRPM is very close to a Gale-Shapley algorithm (Gale and Shapley 1962) that aims at finding stability allocation (stable

matches means that no matched couple would like to break up and forms new matches to be better off). The key aspect from the regulator perspective is that there are several stable sets with different normative properties regarding the side of the market which is favored. Such public algorithms are now implemented by the law in many fields including health care, education (universities and applicants), labor markets, etc.

Cooperative Game Theory and Private Law

Private law – property, torts, and contracts – is also a field studied by cooperative game theory. First, property rights are concerned. In a famous paper on land ownership, Shapley and Shubik (1967) show how different types of property (feudal world, private property, village commune, corporate and joint ownership, etc.) may be modeled in terms of cooperative games. Shapley and Shubik show how the characteristic function has to be changed to well describe different types of property rights.

Second, tort law has been studied through cooperative game theory. Here too, literature is divided in a positive branch and a normative one. From the positive point of view, cooperative game theory has added to a better understanding of the Coase theorem. The Coase theorem is well known in law and economics: in case of zero transaction costs, competition, and perfect delineation of property rights, agents may reach a mutually advantageous agreement. The level of externality is independent of the initial distribution of property rights and maximizes the value of the production (the result is invariant and efficient). Aivazian and Callen (1981) reconsidered the issue and show that the Coasean result is not robust (Coase 1981) when there are more than two players (e.g., several polluters and one victim): the emptiness of the core leads to the impossibility to reach a stable agreement (players have incentive to block any agreement with another coalition). More recently, Aivazian and Callen (2003); Gonzalez et al. (2016); Gonzalez and Marciano (2017), and others provided more complex examples with several victims. The results converge: the Coase theorem does not hold insofar as the delineation of property right influences the emptiness of the core. More importantly, positive transaction costs lead unexpected results: some authors consider that positive transaction costs make the stability of agreement more likely (precisely because renegotiation is costly), and others think not.

In a different perspective, Dehez and Ferey use cooperative game theory from a purely normative point of view as a description of legal adjudication (Ferey and Dehez 2016a, b). The cases studied are tort cases implying multiple tortfeasors who jointly cause a unique harm to a victim or a group of victim. They first consider the Shapley value to estimate the part of the damage to be paid by each tortfeasor, and then they show that the principles of the American Restatement are consistent with the Shapley value principles. Concept solutions are here considered from a normative perspective to better solve conflicts among defendants about their respective shares of responsibility and to describe the behaviors of judges.

In the same vein, a third interesting example regarding private law is about debt and insolvability. In a famous model, Aumann and Maschler (1985) address the issue of the burden of insolvability of a firm. In that case, suppose that players are creditors and get claims against a firm (the debtor) which is unable to pay all its debts back. Suppose E be the total amount of value available (the value of the assets left) and suppose $d1$, $d2$, and $d3$ the claims of the debtors 1, 2, and 3 with $d1 + d2 + d3 > E$. In that case, the worth of each coalition S is defined as the maximum that S can get once all the others creditors ($N\backslash S$) get their claims back, $d(N\backslash S)$. Formally, $v(S) = \text{Max}\ (0,\ d(N\backslash S))$. The authors compare different solution concepts and show that the rules advocated in the Talmud is close to the nucleolus. Their paper is not explicitly oriented in a law and economics perspective, but they suggest that cooperative game theory is very useful to better understand the deep reasons of legal and/or moral reasoning.

Many other examples from private law could be added: patents (several firms cooperate to get a patent), condo (several people have to share the costs of a condo, Crettez and Deloche 2014), or

even the contracts used by bitcoin miners when they pool together to find the relevant hash and have to share the bitcoins get in common. These examples show how fruitful cooperative game theory is for the law.

Conclusion: Cooperative Games and Legal Theory

Law seems to be a quite natural field of application of cooperative game theory. We have provided many examples of legal topics studied with cooperative game theory. To conclude, we would like to add some more speculative views. Is cooperative game theory descriptive, predictive, or normative (Aumann 1985)? Sure, in some cases, cooperative game theory is useful to predict how agents will behave; in other cases, the normative aspect of cooperative game theory is more interesting and provides some axiomatic solutions that help judges or arbitrators to settle conflicts. As Aumann states, "Normative aspects of game theory may be subclassified using various dimensions. One is whether we are advising a single player (or group of players) on how to act best in order to maximize payoff to himself, if necessary at the expense of the other players; and the other is advising society as a whole (or a group of players) of reasonable ways of dividing payoff among themselves. The axis I'm talking about has the strategist (or the lawyer) at one extreme, the arbitrator (or judge) at the other" (Aumann 1985).

Cross-References

- ▶ Coase Theorem
- ▶ Coase Theorem and the Theory of the Core, The
- ▶ REACH Legislation

References

Aivazian VA, Callen JL (1981) The Coase theorem and the empty core. J Law Econ 24:175–181

Aivazian VA, Callen JL (2003) The core, transaction costs, and the Coase theorem. Constit Polit Econ 14:287–299

Aumann RJ (1985) What is game theory trying to accomplish ? In: Arrow K, Honkapohja S (eds) Frontiers of economics. Basil Blackwell, Oxford, pp 5–46

Aumann RJ, Maschler M (1985) Game theoretic analysis of a bankruptcy problem from the Talmud. J Econ Theory 36:195–213

Baird DG, Gertner RH, Picker RC (1994) Game theory and the law. Harvard University Press, Harvard

Beal S, Deschamps M (2016) On compensation schemes for data sharing within the European REACH legislation. Eur J Law Econ 41:157–181

Bittlingmayer (1982) Decreasing average cost and competition: a new look at the *Addyston Pipe* case. J Law Econ 25:201–229

Coase RH (1981) The Coase theorem and the empty core: a comment. J Law Econ 24:183–187

Crettez B, Deloche R (2014) Cost sharing in a condo under law's umbrella. Working paper

Dehez P, Tellone D (2013) Data games: sharing public goods with exclusion. J Public Econ Theory 15:654–673

Ferey S, Dehez P (2016a) Multiple causation, apportionment and the Shapley value. J Leg Stud 45:143–171

Ferey S, Dehez P (2016b) Overdetermined causation, contribution and the Shapley value. Chicago-Kent Law Rev 91:637–658

Gale D, Shapley LS (1962) College admissions and the stability of marriage. Am Math Mon 69:9–15

Gonzalez S, Marciano A (2017) New insights on the Coase theorem and the emptiness of the core. Revue d'Économie Politique 127:579–600

Gonzalez S, Marciano A, Solal P (2016) The Social cost problem, rights and the (non)empty core. Working paper. Available on https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2863952

Lardon A (2017) Coalition games and oligopolies. Revue d'Économie Politique 127:601–636

Littlechild SC, Thomson GF (1977) Aircraft landing fees: A game theory approach. Bell J Econ 8:186–204

Maschler M, Solan E, Zamir S (2013) Game theory. Cambridge University Press, Cambridge

Myerson R (1977) Graphs and cooperation in games. Math Oper Res 2:225–229

Owen G (1977) Values of games with a priori unions. In: Moeschlin O, Hein R (eds) Mathematical economics and game theory: essays in honor of Oskar Morgenstern. Springer, New-York

Roth AE (1984) The evolution of the labor market for medical interns and residents: A case study in game theory. J Political Econ 92:991–1016

Shapley L, Shubik M (1967) Ownership and the production function. Q J Econ 81:88–111

Telser L (1985) Cooperation, competition, and efficiency. J Law Econ 28:271–295

Wiley JS (1987) Antitrust and core theory. J Law Econ 54:556–589

Zhao J (2014) Estimating the merging costs and organizational costs: methodology and the case of 1887–1914 sugar monopoly. Working paper, University of Saskatchewan

Copyright

Antonio Rodriguez Andres[1] and Nora El-Bialy[2]
[1]Departamento de ADE y Economía, Universidad Camilo José Cela, Villanueva de la Cañada, Madrid, Spain
[2]Institute of Law and Economics, University of Hamburg, Hamburg, Germany

Abstract

IPRs indices in general developed to measure the quality or strength of IPR institutions across countries usually do not differentiate between de jure and de facto IPR institutions. In addition, they neglect the different impact these individual variables comprising these indices might have by simply averaging all variables or assigning arbitrary weights. The main contribution of this essay is to shed light on the relative importance of individual institutional variables in forming the de facto institutional framework of IPR protection as opposed to their de jure counterparts mainly focusing on Copyright. For that purpose we discuss the drawbacks of the common indices and offering some suggestions for building more reliable ones. Our main recommendation is to look at the formal copyright institutions in a more careful way to be able to code the different provisions that we think are probably most influential for Institutional Quality and that would enable better enforcement. Second: Countries must make data available about the enforcement process of Copyright laws and number of piracy cases filed in courts and the imposed sanctions in order to be able to develop a de facto measure for the quality of copyright institutions.

Definition

Copyright means individuals are not allowed to copy someone else's works.

Introduction

In legal terms, intellectual property rights (henceforth, IPRs) refer to legal rights to creations of mind like inventions, literary works, artistic works, etc. IPRs are divided into two main categories: industrial property that includes inventions (patents), trademarks, industrial designs, geographic indications of source, copyright, and rights related to copyright. Although the interest in IPRs has been increasing, most empirical research works are focused on patents leaving aside other types of IPR protection. As regards IPR protection, one serious concern for copyright holders is piracy: that is, the unauthorized use of copyrighted goods. Even though piracy occurs for all types of intellectual property and can take several forms depending on the access type and intellectual property mechanism, one of the most worrying areas is the piracy of business software applications. This entry addresses the study of copyright institutions and intends to measure the strength of the IP systems for copyright institutions for a cross-national sample as other authors have done for copyright laws in Europe (Andres 2006).

The divergence between institutional reforms (legal reforms) and factual implementation of copyright laws has recently become more critical and apparent (Table 1 in the appendix shows the number of IPR laws issued or amended and related agreements signed by each country as opposed to the prevailing software violations in order to show that the existence of copyright institutions is necessary but cannot be considered a sufficient condition for enforcement). The correlation between the number of copyright laws issued and the perceived level of IPR protection is −0.34. In the absence of enforcement and adequate sanctions, IPR reforms and signed agreements tend to fall short of their proclaimed goals, leading to inefficient IPR institutions. Hence, in order to design an efficient IPR reform strategy, one must first identify whether the failure of existing IPR laws is due to the lack of formal written aspects of IPR legislations (de jure IPR institutions) or it is caused by the inefficiency of enforcement authorities responsible for implementing the law (de facto IPR institutions).

Copyright, Table 1 The number of IPR-related laws issued and agreements signed by each country

	Agreements	Laws	Software piracy rates
Argentina	14	7	69.88
Armenia	9	3	92.33
Australia	15	14	30.52
Austria	15	14	32.52
Azerbaijan	10	1	91
Belgium	12	8	32.70
Bolivia	11	12	83.05
Bosnia	15	3	67.71
Brazil	14	5	61.47
Bulgaria	9	1	77.64
Canada	8	15	36.88
Chile	11	14	60.65
China	9	19	89.29
Colombia	9	11	58.52
Costa Rica	15	5	69
Croatia	13	7	64.1
Cyprus	12	2	59.76
Czech Republic	13	10	44.47
Denmark	15	2	29.70
Ecuador	15	13	70.76
Egypt	5	2	68.59
El Salvador	8	5	82.82
Estonia	7	11	52.2
Finland	12	9	31.70
France	16	34	44
Germany	17	15	31.23
Greece	12	16	67.29
Honduras	10	5	75.58
Hong Kong	1	16	55.17
Hungary	13	3	50.88
Iceland	6	23	50.66
India	12	5	70
Indonesia	5	2	89.23
Ireland	9	30	47.2
Israel	9	12	43.9
Italy	16	38	50.2
Japan	14	21	32.9
Jordan	6	4	69.3
Kazakhstan	7	0	80.9
Kuwait	2	1	74.5
Latvia	9	2	56.9
Lithuania	9	10	55.5
Luxembourg	11	4	32.8
Malaysia	4	10	66.6
Malta	6	4	54.6
Mauritius	8	2	67.7
Mexico	13	11	62.1
Morocco	13	9	68.9
Netherlands	14	9	39.2
New Zealand	9	17	27.7
Nicaragua	14	9	81.6
Norway	12	14	36.6
Pakistan	7	3	85.7
Panama	16	7	70.2
Paraguay	13	2	83.2
Peru	14	14	69.6
Philippines	11	15	74.2
Poland	11	4	59.9
Portugal	15	31	45.8
Romania	7	3	76.2
Russia	12	21	83.8
Saudi Arabia	7	4	60.1
Serbia	15	2	76.6
Singapore	7	11	46.3
Slovakia	14	4	49.2
Slovenia	17	9	63
South Africa	5	12	42
Spain	13	16	51.8
Sweden	16	10	33.6
Switzerland	30	6	32
Thailand	4	3	79.4
Tunisia	11	3	77.4
Turkey	7	7	70.8
Ukraine	11	13	88.9
United Kingdom	15	42	28.9
United States	15	7	23.4
Uruguay	13	16	71
Venezuela	10	11	72.6
Vietnam	7	14	92.8
Zambia	4	1	82.2

Sources: Software piracy rates. Business Software Alliance (BSA) (2009) Sixth annual BSA and IDC global software piracy studies (online). BSA. http://global.bsa.org/globalpiracy2008/index.html. Accessed Mar 2010

The first option requires a reform strategy targeting improvement of the existing IPR laws and regulation by issuing, for example, further enforcement or sanctions. The latter requires

C

launching a judicial reform or improving the quality of the enforcement organizations in general.

It is worth noting that most pirating or counterfeiting countries have already issued national IPR laws and may even have special provisions that serve as a copyright law with a relevant sanction mechanism. However, these laws have not translated into significantly lower software piracy rates. Hence, it appears that the existing IPR institutions might turn out to be inefficient, especially in most developing countries. In such cases, the use of IPR rules and legal provisions might be an unreliable measure for reflecting the level of IPR enforcement in a country. Thus, there is a wide gap in our understanding of how to measure the quality of copyright institutions. Therefore, measuring the strengths of IP systems for copyright at country level may enable a better policy design by giving information about the determinants of IP protection, as well as their economic and social effects. This is the problem this entry addresses.

Measuring the Quality of Copyright Institutions

The level of economic development is considered a main determinant of all forms of IPR piracy including copyright infringement across countries (Maskus and Penubarti 1995; Park and Ginarte 1997; Gopal and Sanders 1998, 2000; Maskus 2000). With the recent rise of the new institutional economics (NIE) and the growing role of institutions in economic studies, scholars started to highlight the impact of institutions on IPR piracy.

Institutions in general comprise a rule that is accompanied by a sanction. Accordingly, institutions that lack a sufficient sanctioning mechanism cannot fulfill their role and end up written down *only* in formal decrees. Hence, setting institutions that are in charge of administrating the system constitutes just one of the pieces that form a national system of IPR protection. These institutions must interact with some organizations in order to realize factual implementation and protect these rights and enforce the institutions. North (1981) described the process of property right enforcement as making it obligatory or put it into force. This process protects right holders by preventing the infringement of property rights by free riders. Thus, organizations in charge of enforcing rights become crucial elements of the system's overall effectiveness, as they play a main role in setting the legal infrastructure.

Institutions are just the rules of the game, while organizations are the players. Hence, it can be said that the efficiency of institutions and the performance of organizations are highly interdependent.

Mansfield (1994) determined three areas of concern in assessing the strength of IPR protection in a country. These are (a) existing laws, (b) the prevailing legal infrastructure, and (c) the willingness of governments to actively enforce property rights. Whereas development economists have typically treated the state as an exogenous actor in the development process, North (1990), however, pointed out that the state can never be treated as an exogenous actor in development policy. He explained that third-party enforcement represents the development of the state as a coercive force able to monitor property rights and enforce contracts effectively. In other words, this third party is responsible for enforcing the contracts carried out by different parties and punishing the party that violates the agreement (Harris et al. 1995) (for further information, see Harris et al. 1995, p. 22; Sened 1997, p. 49). The enforcement of agreements in economic markets is ultimately a function of the political markets of economies. The political market imposes the polity that specifies property rights and provides the instruments and resources to enforce contracts. Hence, an explanation of property right enforcement in general requires an understanding of the behavior of the government and assumption that government is exogenous because government normally defines and enforces property rights (North 1990). Even so, the government is not the sole actor of the enforcement process, as its main role practically ends after formally signing the law. Afterward, the implementation process will be handed to the law enforcers which are responsible for deterring and sanctioning those who break the law. It starts with the police who raid the suspected sites and catch infringers to file a case. After filing it, the case will be handed over

to the prosecutors to gather all the needed information and evidence to finally submit it to the courts. Hence, it can be considered that the overall efficiency of the formal enforcement system in a country to handle copyright infringement among other cases of commercial or civil trials plays a huge role in determining the adequacy of copyright law enforcement. This process, where it is inadequate, requires special reforms within the enforcement organizations themselves in order for the participating bodies (police, prosecutors, and judiciary) to enjoy a high degree of accountability. Williamson (1975) analyzes the overall impact of poor performance of courts on contractual obligations. The study shows that court litigations are time-consuming and might result in errors or in the absence of choice. This can happen when judges dismiss cases due to wrong litigation procedure or the lack of sufficient evidence, which is typical for most copyright suits.

Aspects of the "De Jure" Copyright Institutions

De jure IPR institutions include formal institutions (laws, formal amendments of rules, agreements, etc.) designed to protect IPRs.

For research purposes, many indicators are used to measure the quality of institutions within the IPR research context. For instance, using a measure of institutional strength developed by Knack and Keefer (1995) as an explanatory variable for IPR infringement, Marron and Steel (2000) argue that institutional factors in general have greater influence on piracy than economic conditions of a country. This indicator uses data published in the firm's International Country Risk Guide (ICGR) and identifies five variables related to property rights protection, which are tradition of law and order, the government's propensity to repudiate contracts, the quality of bureaucracy, the extent of corruption, and risk of expropriation. By contrast, most studies use either a dummy variable to represent the strength of IPR institutions in a country or an index composed of multiple variables reflecting the quality of institutions at the national level. Both methods are problematic to some extent as the brief discussion below illustrates.

A Dummy Variable Approach

Using a dummy variable representing a country's membership in international agreements or the possession of a national IPR code is often used to represent copyright institutions (e.g., Papadopoulos 2003; Van Kranenburg and Hogenbirk 2005; Andres 2006). However, as mentioned before, most countries are already members in one or more international IPR agreements and already have issued their own IPR codes. Hence, the use of a dummy variable to represent membership in agreements or IPR laws would be inappropriate. This is a necessary condition but not sufficient as argued by Maskus (2000). As argued in the theoretical part of the study and supported by the data observed from reviewing the IPR laws and regulations of different countries, it is noticed that countries keep changing their IPR laws by adding extra provisions, extending the scope of coverage, or maybe even issuing a new law to ensure a better enforcement. For this purpose, counting the number of laws issued or agreements signed by a country over time might be considered as a proxy for measuring the efforts made toward improving IPR institutions rather than reflecting the qualities of these institutions. The existence of the copyright institution per se cannot be considered a determinant of its quality, but rather we should focus on the details and the provisions included in the law and the efforts of the legislators to improve the code by doing amendments and modifying the framework of the law to suit the countries' conditions. Especially, in copyright software piracy matters and with the fast growing technology and innovative piracy devices, codes and provisions have to be updated through time to ensure clear enforcement provisions.

Another appropriate measure would be constructing variables through revising existing copyright laws and international copyright agreements in order to search for differences among the single provisions of the different laws. The index for patent rights developed by Ginarte and Park (1997) was the first attempt toward coding the content and scope of the laws but with the fast developments of the IP

codes in most countries. They have composed a patent protection index for a panel of 110 countries from 1960 to 1990. It uses an unweighted sum of five variables to build up the index. The variables are the scope of coverage, membership in international agreements, provisions for loss protection, enforcement mechanisms, and the duration of IPR protection (for more details about the composition of the index, see Ginarte and Park (1997), pp. 286–287). Each variable is assigned a value of one if present and zero otherwise, and then all of them are aggregated to obtain a maximum score of 5.0 in case all 5 factors were taken care of in each countries' law. However, there have been a variety of new laws issued and agreements signed, such that the differences in the variables used in their index became invariant among countries to a large extent. One important remark is that the indexes are largely measuring the laws on the books rather the actual practice. A common concern is that these IP indexes measure only the perceived protection not actual protection. As argued by Park (2001), nevertheless, there is a high correlation between statutory laws and the current level of enforcement. Countries that have strong laws on the books tend to be the ones that carry out the laws. Hence, developing measures that capture the significantly different aspects of IPR codes across countries might be a more proper way to measure IPR institutions (see the UNESCO portal: http://portal.unesco.org/culture/en/ev.php- URL_ID = 39069&URL_DO = DO_TOPIC&URL_SECTION = 201.html and WIPO: http://www.wipo.int/clea/en/ and http://www.wipo.int/about-ip/en/ipworldwide/country.htm). For example, one could code the severity of sanction provisions if mentioned at all in a law. In addition, one could also count the number of clauses within each country law regarding the implementation and enforcement procedure of a piracy crime. It might be expected that countries having specified implementation clauses in their law will observe better law enforcement and hence experience lower piracy rates. Finally, a variable measuring the number of copyright laws and regulations issued by executives rather than the legislative body might also be significant in determining the quality of copyright institutions.

Aspects of the "De Facto" Copyright Institutions

In an attempt to measure the factual IPR law implementation or determine the quality of enforcement, formal empirical studies used the level of corruption in a country as a proxy for institutional quality (e.g., Ronkainen and Guerrero-Cusumano 2001; Papadopoulos 2003; Bagchi et al. 2006). The indicators of Kaufmann et al. (2011) "graft/corruption index" (Kaufmann et al. 2011) in addition to the Transparency International's 1999 Corruption Perception Index (CPI) (Lambsdorf 2000) are mainly used in the previous studies. These indices reflect the compilation of perceptions of the quality of governance and are used to represent the overall corruption level among custom agents, police, prosecutors, and judicial in the country.

Whereas Shadlen et al. (2005) employ the Kaufmann et al. "government effectiveness indicator" to examine the effect of institutional factors on piracy, Holm (2003) considers the "rule of law" as most appropriate when attempting to proxy for the institutional aspects of a country, as it aims to measure the efficiency of the judicial system (available in Kaufmann et al. (2004)). Fischer and Andrés (2005) explain that the degree of efficient law enforcement which is determined by the rule of law variable may increase the probability of imposing a deterrent punishment; hence, it may best capture the de facto aspects of copyright institutions. The rule of law measures the government's administrative capacity in enforcing the law. More recently, Ping (2010) uses the rule of law to measure laws and institutions and finds it to be an important determinant of the rate of software piracy.

However, using the Worldwide Governance Indicators (henceforth, WGI) of Kaufmann et al. (2006, 2011) as measures of the institutional quality of a country might be however somehow problematic (the WGI comprise six dimensions of governance: voice and accountability, political stability and absence of violence, government effectiveness, regulatory quality, rule of law, and

control of corruption). In addition, most of these bundled indicators may be criticized by the fact that they do not focus on a single variable or institution but relate to dozens or even hundreds of variables that are not equally weighted and are not necessarily related to each other. Also note further that the weights of individual variables appear to be different when comparing the old with new WGI versions. Voigt (2013), e.g., shows that most of the partial correlations between a selected number of the rule of law variables are quite low, which implies that the rule of law is indeed made up of various dimensions that are not necessarily highly correlated with each other. Moreover, each variable or group of variables synthesized in each indicator is weighted differently from another indicator. Hence, it can be said that the indicator does not represent the variables with equal strength, which may hinder one's ability to identify the relative significance, if any, of each institution or variable individually.

We argue that it might be useful to unbundle the indices and to consider the performance of the different enforcement organizations (police, judiciary, executives) individually to measure de facto IPR institutions in a more accurate way. For that purpose, we raise the following two questions: (i) Is the IPR enforcement process treated in the way a regular penal act is treated (pirates are caught by the police), or is it handled by "specialized agencies? (ii) Are copyright infringement cases filed in specialized courts, or are they counted as civil or criminal acts and hence treated in regular first instance courts? All these details matter; if copyright infringement cases are treated in a special manner and in a way that deviates from the regular penal cases, then accounting for the quality or accountability of the police and the regular court becomes irrelevant and hence would not be a proper measure for de facto copyright institutions and hence enforcement.

Cross-References

► Economic Analysis of Law

Appendix

See Table 1.

References

Andrés A.R. 2006. The relationship between copyright software protection and piracy: evidence from europe. European Journal of Law and Economics, 21, 29–51

Bagchi K, Kirs P, Cerveny R (2006) Global software piracy: can economic factors alone explain the trend? Commun ACM 49(6):70–75

Business Software Alliance (BSA) (2009) Sixth annual BSA and IDC global software piracy studies (online). BSA. http://global.bsa.org/globalpiracy2008/index.html. Accessed Mar 2010

Fischer J, Andrés A (2005) Is software piracy a middle class crime? Investigating the inequality-piracy channel. University of St. Gallen working paper series, Department of Economics, University of St. Gallen

Ginarte JA, Park W (1997) Determinants of patent rights: a cross-national study. Res Policy 26:283–301

Gopal R, Sanders G (1998) International software piracy: an analysis of key issues and impacts. Inf Syst Res 9(4):380–397

Gopal RA, Sanders G (2000) Global software piracy: you can't get blood out of a turnip. Commun ACM 43(9):82–89

Harris J, Hunter J, Lewis C (1995) The new institutional economics and third world development. Routledge, London

Holm H (2003) Can economic theory explain piracy behaviour? Top Econ Analy Pol 3(5):1–15

Kaufmann D, Kraay A, Mastruzzi M (2006) Governance matters V: governance indicators for 1996–2005. World Bank policy research working paper

Kaufmann D, Kraay A, Mastruzzi M (2011) Worldwide governance indicators Acessed at: http://info.worldbank.org/governance/wgi/index.asp, Accessed on 20 Feb 2014

Knack S, Keefer P (1995) Institutions and economic performance: cross-country tests using alternative institutional measures. Econ Polit 7:207–227

Mansfield E (1994) Intellectual property protection, foreign direct investment and technology transfer, International finance corporation discussion paper, 19. The World Bank. http://www.bvindecopi.gob.pe/colec/emansfield2.pdf. Accessed Oct 2008

Marron D, Steel D (2000) Which countries protect intellectual property? The case of software piracy. Econ Inq 38:159–174

Maskus K (2000) Intellectual property rights in the global economy. Institute for International Economics, Washington, DC

Maskus K, Penubarti M (1995) How trade related are intellectual property rights? J Int Econ 39: 227–248

North D (1981) Structure and change in economic history. Norton, New York
North D (1990) Institutions. Institutional change and economic performance. Cambridge University Press, New York
Papadopoulos T (2003) Determinants of international sound recording piracy. Econ Bull 6(10):1–9
Png IP (2010) On the reliability of software piracy statistics. Electronic Commerce Research and Applications 9:366–373.
Park W (2001) Intellectual property and economic freedom. In J. Gwartney and R. Lawson (eds.), Economic Freedom of the World, Vancouver: Fraser Institute
Park W, Ginarte J (1997) Intellectual property rights and economic growth. Contemp Econ Policy 15(3):51–61
Ronkainen I, Guerrero-Cusumano J (2001) Correlates of intellectual property violation. Multinatl Bus Rev 9(1):59–65
Sened, I (1997) The political institution of private property. Cambridge: Cambridge University Press
Shadlen K, Schrank A, Kurtz M (2005) The political economy of intellectual property protection: the case of software. Int Stud Q 49:45–71
Van Kranenburg H, Hogenbirk A (2005) Multimedia, entertainment, and business software copyright piracy: a cross-national study. The Journal of Media Economics, 18:109–129
Voigt S (2013) How (not) to measure institutions? J Institut Econ 9(1):1–26
Williamson O (1975) Markets and hierarchies: analysis and antitrust implications: a Study in the economics of internal organization. University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship (online). http://ssrn.com/abstract=1496220. Accessed Mar 2010
WIPO (2001) WIPO intellectual property handbook: policy, law, and use, vol 489(E), WIPO Publication. WIPO, Geneva
Lambsdorf J (2000) Background paper to the 2000 corruption perceptions index, Transparency International, Sept. www.transparency.org. Accessed on Feb 2014

Corporate Criminal Liability

Paolo Polidori and Désirée Teobaldelli
Department of Law, University of Urbino, Urbino, Italy

Abstract

This entry reviews the literature on corporate criminal liability. It first describes the different forms of corporate liability and then discusses the optimal structure of corporate sanctions to deter crimes. The distinction between civil and criminal corporate liability is addressed, and a brief discussion of the corporate criminal enforcement in the United States and Europe is presented.

Definition and Structures of Corporate Liability

Corporate liability is the liability of one party (the firm) for the misconduct of another party (the employee). Corporate liability can assume different forms ranging from *vicarious liability* to some forms of *duty-based liability*. Vicarious liability is a strict (absolute) form of secondary liability that arises under the legal doctrine of *respondeat superior*, that is the common law doctrine of agency for which a party (the principal) is responsible for the acts committed (within the scope of employment) by its agents that has the legal right or duty to control. Duty-based liability regimes are forms of secondary liability where the principal is held liable for the misconduct of the agent only if he contravenes a legal duty – that is, if he does not observe due care in preventing or reporting a violation. There may also be mixed regimes under which firms are liable, but the level of sanctions depends on whether the corporation fulfilled its policy responsibilities.

Besides being vicarious or duty-based, corporate liability may be civil or criminal. This entry addresses the theoretical reasons for corporate criminal liability by discussing the following three issues:

(i) Under what conditions it is optimal to impose sanctions also on the firm (principal) rather than only on the employee (agents) who misbehaved, namely, why there should be *joint* individual and corporate liability?
(ii) Which form of corporate liability, vicarious versus duty-based, is socially optimal?
(iii) Whether the socially optimal corporate liability should be criminal or not. We then briefly discuss the corporate liability in the USA and Europe.

The Need of Joint Individual and Corporate Liability

To understand the optimal deterrence of corporate crimes, it is useful to keep in mind that such crimes are crimes committed by individuals working for firms, i.e., they are crimes committed in presence of an agency relationship between the firm and the employee. However, the starting point of the analysis is represented by the results of the classic model of individual criminal liability in absence of corporations when individuals are risk neutral and are not wealth constrained and sanctioning costs are zero (Becker 1968; Polinsky and Shavell 2000). As individuals will commit a crime only when the expected benefit exceeds the expected cost, the socially optimal level of deterrence requires that the state imposes a criminal fine equal to the ratio between the social cost of crime to society and the probability of crime being detected. This result also holds when the probability of detection depends on the enforcement expenditure with the caveat that the optimal deterrence scheme involves minimizing enforcement costs, which means that the probability of crime detection is at its lower bound. The analysis for unintentional crimes leads basically to same results as the state will induce the optimal individual's level of effort to avoid the crime by choosing a sanction scheme that equalizes the expected sanction to the social cost of crime.

Once we introduce corporations into the analysis, the first issue to be addressed is whether corporate liability is necessary to optimally prevent corporate crimes. Let us begin considering a framework without imperfections, i.e., a world where the following conditions are satisfied:

1. Firms and employees have no wealth constraints.
2. There is a positive probability that the state sanctions a crime even without spending resources on enforcement.
3. There are no costs on imposing sanctions.
4. All parties are rational and there is perfect information.
5. Firms and employees can contract at zero cost.

In this perfect world, there is no justification for joint individual and corporate liability because crime can be optimally deterred by imposing the liability on either the individual or the firm at the same social and private costs. As the state is indifferent between individual and corporate liability (i.e., the two forms of liability are complete substitute in this framework), this result is known in the literature as the neutrality principle (Kornhauser 1982; Sykes 1984; Polinsky and Shavell 1993).

The assumptions on which the neutrality principle is based are very strong and are generally not satisfied in reality. This opens the way for an important role played by corporate liability. In particular, it is immediate that the state cannot optimally deter corporate crime using individual liability and monetary sanctions because employees do not generally have sufficient assets to pay the sanctions; corporate crimes often generate large social costs (see, e.g., the case of financial and consumer frauds), and the probability of sanction for such kind of crimes tends to be low, especially in absence of significant expenditures on enforcement. Again, while imprisonment can increase the level of sanctions imposed on the wealth constrained individuals, this is unlikely to make individual liability alone sufficient to deter corporate crime for the following reasons:

(a) The maximal punishment, i.e., life imprisonment, is likely to imply a limited duration of this penalty given that corporate wrongdoers tend to be relatively old.
(b) Imposing long prison sentences for nonviolent crimes, such as corporate crimes, may not be possible because this reduces the room of appropriate sanctions for violent criminal offences.
(c) Imprisonment entails very large costs to society both because of the standard reasons associated to the detention of individuals (e.g., removal of productive individuals from the labor market, expenditures for guards, protective services, and equipment) and the losses suffered by risk-averse agents who face the risk of being liable also when they have not committed the crime.

Another important reason behind the failure of the neutrality principle and the need of corporate liability is related to the complexity of corporate crime as such crimes may involve many individuals and determining those responsible for the wrongdoings is often a difficult task. This means that optimal deterrence requires that both the state and the firm spend resources on prevention and enforcement. In particular, the firm can play an important role in deterring corporate crimes by:

(i) Adopting measures of crime prevention that reduce the incentives to commit a crime; this can be made by adopting policies, such as those related to compensation and promotion that influence the extent to which the employee benefits from the crime (e.g., firms may limit the adoption of high-powered short-run compensation policies who generally provide an incentive to managers to commit crimes; Arlen and Carney 1992); corporations can also make committing crime more costly by promoting a culture of legal compliance within the firm so to increase the likelihood that employees report suspect wrongdoings or that wrongdoers bear higher direct psychological costs (Conley and O'Barr 1997; Tyler and Blader 2005).
(ii) Implementing policing measures that increase the probability of crime detection and conviction; this can be done ex ante through the adoption of compliance (monitoring) programs that allows the firm to detect crime and to collect evidence on wrongdoers more easily and ex post (i.e., after the crime is committed) by reporting detected crimes and by cooperating with the authorities to investigate the crime (Arlen 1994; Arlen and Kraakman 1997).

In sum, corporate crimes are committed in presence of an agency relationship between the firm and the employee and optimal deterrence needs to take into account that the firm has a comparative advantage in the collection of information relative to the state. Therefore, optimal deterrence of corporate crimes requires that the state not only has to invest optimally in enforcement and to impose the optimal sanctions on individuals but also has to provide the optimal incentives to firms to undertake policies of prevention and policies. In other words, corporate liability is essential (Arlen and Kraakman 1997; Arlen 2012). Of course, there are many factors affecting the decision of the firm to prevent crimes for a given structure of corporate liability; for example, variables such as the firm's size, its productivity, and its market power as well as the profitability of illegal activities end up being important determinants of whether (and how) the firm prevents manager wrongdoings (Polidori and Teobaldelli 2016).

Individual liability in presence of corporate liability, and therefore the joint liability, is necessary because under various circumstances, the firm may be unable (or find it not optimal) to impose the optimal level of sanctions on employees. One of these situations, especially relevant for closely held firms and smaller publicly held firms, is when the firm has insufficient asset to pay the optimal corporate sanction. In this case the firm may find it optimal to impose a suboptimal level of sanctions to the employees (Kornhauser 1982; Kraakman 1984; Sykes 1984); moreover, pure corporate liability is likely to create other kind of distortions as managers have the incentive to keep the firm undercapitalized. In larger publicly held firms, individual liability is necessary because of the agency problems characterizing these firms where there is a separation of ownership and control. Large corporations may be unable to impose the optimal level on sanctions to their employee because managers can often influence the decisions of the board of directors, directly or by controlling the information available to the board (Arlen and Carney 1992). In all these cases, individual liability ensures that the state can impose larger (monetary and nonmonetary) sanctions than the firm to the employee responsible of the wrongdoer and improve social welfare (Segerson and Tietenberg 1992; Polinsky and Shavell 1993).

Vicarious Versus Duty-Based Liability

The optimal structure of corporate liability requires that the firm deters crime by adopting measures of crime prevention and by undertaking policing activities at the optimal level. This result cannot be obtained under vicarious liability because the activities of prevention and policing generate a liability enhancement effect, namely, an increase of the firm's expected liability for the crimes that the firm does not deter (or that cannot be deterred), which reduces the firm's incentive to pursue such polices. Indeed, under strict liability, the activities of crime prevention (such as the adoption of monitoring programs) increase the probability that crimes will be detected and the firm sanctioned (Arlen 1994); similarly, the ex post policing (self-reporting and cooperation with the authorities) may reduce crime ex ante, but it increases the probability that committed crimes are sanctioned so that the net effect on the firm's expected sanctions may well be positive (Arlen and Kraakman 1997).

While strict corporate liability may provide to the firm too little incentive for measures of prevention and policing, a multitiered duty-based sanction regime may overcome this problem and induce the firm to monitor and police optimally (Arlen 2012). More precisely, the state should induce the firm to:

(i) Employ crime-preventing measures by imposing a penalty if it has not monitored optimally (e.g., it has not adopted appropriate monitoring programs); the penalty can be avoided by the firm choosing the optimal monitoring.
(ii) Engage in ex post policing by imposing additional sanctions that can be avoided if the firm self-report the crime and fully cooperate with the authorities to convict the wrongdoers.

Although some authors (e.g., Weissmann 2007) argue that the firm should not be liable if it prevents and police optimally, others argue that the optimal structure of corporate liability also requires that the state imposes a residual strict corporate liability. This residual liability (that should be civil, not criminal) is required to eventually impose additional sanctions equalizing the firm's expected sanction, net of the costs beard by the firm through market sanction, to the total social cost of crime (Polinsky and Shavell 1993; Shavell 1997; Arlen and Kraakman 1997). The main market sanction is generally represented by the reputational penalties which reflect the greater difficulty of *criminal* firms in contracting with other parties on favorable terms; the market's anticipation that the firm will produce lower profits – owing to either higher costs or lower revenues – may lead to a reduction in the firm's value (Klein and Leffler 1981; Karpoff and Lott 1993).

The works on reputational penalties suggest there are large variations in the size of these penalties depending on the type of crime. In particular, parties contracting with the firm will react negatively to news that the firm committed crimes if those crimes (such as fraud) harm them or other contracting parties; however, theory also suggests that firms should not be punished by the market for crimes committed by noncontracting third parties as may occur in cases of regulatory or environmental violation (Karpoff and Lott 1993; Alexander 1999). The available empirical evidence provides support to these theoretical arguments (Karpoff et al. 2005, 2017). Higher reputational penalties clearly provide an incentive to the firm to prevent crime, and this effect is likely to be larger when the firm operates in more competitive markets (Polidori and Teobaldelli 2016).

Civil Versus Criminal Corporate Liability

The doctrine of corporate criminal liability is controversial for various reasons. One of these reasons is the conceptual difficulty of imposing criminal liability on a corporation, an artificial and collective entity, given that criminal liability requires a culpable mental state or *mens rea* (Alexander 1999; Hamdani and Klement 2008). At the same time, one of the main advantages of the criminal law system, that is, the sanction of incarceration as a punishment, is unavailable

against enterprises as these are not physical entities and, therefore, cannot be imprisoned (Khanna 1996; Fischel and Sykes 1996). This consideration leads to the second critical issue concerning the rationale for imposing criminal liability on a corporation given that the criminal and civil corporate sanctions are quite similar in nature (as they are mainly monetary) and can have the same magnitude.

Despite the similarity between corporate criminal and civil liability, they differ along some dimensions. First, there are important procedural differences; criminal cases require a higher burden of proof for the conviction of the wrongdoers, but the authorities have more powerful tools of investigation. Second, whereas corporate criminal and civil liability share the form of monetary sanctions, those levied for criminal behavior are much stronger and can lead to enormous collateral sanctions (e.g., debarment and de-licensing); furthermore, criminal sanctions are not only higher than civil penalties but can also be imposed in addition to them. Third, criminal sanctions are generally associated with much larger reputational losses than civil penalties.

The key distinctive feature of criminal liability of imposing larger sanctions on convicted firms may allow the state to structure a multitiered duty-based liability that induces the firm to prevent and police optimally more efficiently than a pure civil liability regime. However, criminal penalties may also create the risk of overdeterrence of crimes which in case of corporate crimes may have important negative effects on social welfare. Therefore, a well-structured duty-based liability is essential to produce socially efficient outcomes.

Corporate Criminal Liability in the United States and in Europe

Corporate criminal enforcement in the United States is characterized by joint individual and corporate criminal liability for business crimes and firms are formally subject to a strict de jure liability regime for employees' crimes. However, criminal liability is de facto duty-based as firms can avoid indictment if they self-report wrongdoing and cooperate with government authorities to convict individual wrongdoers. This Department of Justice policy was initiated in 1999 by Eric Holder (then the US Deputy Attorney General) who issued guidelines to federal prosecutors on when firms should be indicted for employees' crimes committed during the scope of their employment. The Holder memo encouraged prosecutors not to indict firms for employee crimes if they adopted compliance programs, self-reported the wrongdoings, and fully cooperated with the federal authorities. In the existing regime firms that report and cooperate are still subject to some form of expected monetary sanction; prosecutors impose both monetary and nonmonetary sanctions by means of deferred prosecution and non-prosecution agreements (DPAs and NPAs, respectively). These agreements often impose also nonmonetary performance mandates on the firm; for example, they may require the firm to adopt prosecutor-approved compliance programs or to appoint an outside corporate monitor who reports to federal authorities (Garrett 2007; Arlen and Kahan 2017). Arlen (2012, p. 152) concludes that US enforcement practice is consistent with the optimal level of corporate liability.

Even though formal conditions governing leniency are quite broad and could apply to all firms, large and publicly held firms typically avoid formal conviction (by way of DPAs and NPAs), whereas small owner-managed, closely held firms are usually convicted (Arlen 2012). The most likely reason for this difference is that cooperation with prosecutors is the key element determining whether prosecution is avoided or not, and closely held firms tend to cooperate less because they are inclined to protect their owner/managers. In contrast, in large publicly held firms, the owners are rarely directly involved in day-to-day management, and decisions to cooperate with the authorities are often delegated to outside directors who have strong incentives to cooperate in return for leniency.

At the European Union level, after a number of harmonization efforts, there is a general trend in almost all the Member States toward the adoption of corporate criminal liability regimes. With the

exception of the UK and the Netherlands, where there is a long history of recognizing corporate criminal liability, the concept of corporate criminal liability is relatively recent in Europe. The first European country that introduced corporate criminal liability was France in 1994, followed by Finland in 1995, Denmark in 1996, Belgium in 1999, Italy in 2001, Poland in 2003, Austria in 2005, Romania in 2006, Portugal in 2007, Luxembourg and Spain in 2010, Czech Republic in 2012, and Slovakia in 2016. Germany represents a remarkable exception, since corporate criminal liability is considered incompatible with the essence of German criminal law based on the notion of individual culpability; however, in late 2013, the Department of Justice of North Rhine-Westphalia presented a first draft of a Corporate Penal Code, although this remained under discussion (Clifford Chance 2016).

Despite harmonization efforts, relevant differences still exist between the Member States, due to different legal tradition and, also, to different approaches followed in the implementation of the proposed EU legislation. While most countries applied criminal procedure rules, others have imposed quasi-criminal administrative liability regimes. In some jurisdictions, the category of employees which can activate corporate liability is restricted to those with management responsibilities, and the employee's misconduct must be done in the interests of or for the benefit of the company. Similarly to the USA, also at the EU level, a common feature characterizing the legal framework is the possibility for the corporation to have a reduction of the potential penalties in case of adoption of adequate compliance systems to prevent the crime and of cooperation with the authorities. While penalties vary across countries, many of them complement the standard monetary sanctions with nonmonetary sanctions such as judicial supervision of a corporation's affairs, exclusion from public procurement tenders, limitations on the use of checks and credit, confiscation of assets gained from the offence, posting of notices throughout the media, publication of the judgment in a registry, and even the dissolution of the corporation, in the most severe cases (Sun Beale and Safwat 2004).

References

Alexander CR (1999) On the nature of reputational penalty for corporate crime: evidence. J Law Econ 42:489–526

Arlen JH (1994) The potentially perverse effects of corporate criminal liability. J Leg Stud 23(2):833–867

Arlen JH (2012) Corporate criminal liability. Theory and evidence. In: Harel A, Hylton KN (eds) Research handbook on the economics of criminal law. Edward Elgar, Northampton, Massachusetts, USA

Arlen JH, Carney WJ (1992) Vicarious liability for fraud on securities markets: theory and evidence. Univ Ill Law Rev 1992:691–740

Arlen JH, Kahan M (2017) Corporate governance regulation through nonprosecution. Univ Chic Law Rev 84 (1):323–387

Arlen JH, Kraakman R (1997) Controlling corporate misconduct: an analysis of corporate liability regimes. N Y Univ Law Rev 72(4):687–779

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217

Clifford Chance (2016) Corporate criminal liability. Available at www.cliffordchance.com

Conley JM, O'Barr WM (1997) Crimes and custom in corporate society: a cultural perspective on corporate misconduct. Law Contemp Probl 60:5–22

Fischel DR, Sykes AO (1996) Corporate crime. J Leg Stud 25:319–349

Garrett BL (2007) Structural reform prosecution. Va Law Rev 93(4):853–957

Hamdani A, Klement A (2008) Corporate crime and deterrence. Stanford Law Rev 61:271–310

Karpoff JM, Lott JR Jr (1993) The reputational penalty firms bear from committing fraud. J Law Econ 36(2):757–802

Karpoff JM, Lott JR Jr, Wehrly EW (2005) The reputational penalties for environmental violations: empirical evidence. J Law Econ 48(2):653–675

Karpoff JM, Lee DS, Martin GS (2017) Foreign Bribery: incentives and enforcement. Available at SSRN: https://ssrn.com/abstract=1573222

Khanna VS (1996) Corporate criminal liability: what purpose does it serve? Harv Law Rev 109:1477–1534

Klein B, Leffler KB (1981) The role of market forces in assuring contractual performance. J Polit Econ 89(4):615–641

Kornhauser LA (1982) An Economic analysis of the choice between enterprise and personal liability for accidents. Calif Law Rev 70(6):1345–1390

Kraakman RH (1984) Corporate liability strategies and the costs of legal controls. Yale Law J 93(5):857–898

Polidori P, Teobaldelli D (2016) Corporate criminal liability and optimal firm behavior: internal monitoring versus managerial incentives. Eur J Law Econ 1–34. https://doi.org/10.1007/s10657-016-9527-2

Polinsky AM, Shavell S (1993) Should employee be subject to fines and imprisonment given the existence of corporate liability. Int Rev Law Econ 13:239–257

Polinsky AM, Shavell S (2000) The economic theory of public enforcement law. J Econ Lit 38(1):45–76

Segerson K, Tietenberg T (1992) The structure of penalties in environmental enforcement: an economic analysis. J Environ Econ Manag 23:179–200
Shavell S (1997) The optimal level of corporate liability given the limited ability of corporations to penalize their employees. Int Rev Law Econ 17:203–213
Sun Beale S, Safwat A (2004) What developments in Western Europe tell us about American critiques of corporate criminal liability. Buffalo Crim Law Rev 8:89–163
Sykes AO (1984) The economics of vicarious liability. Yale Law J 93(7):1231–1280
Tyler TR, Blader SL (2005) Can businesses effectively regulate employee conduct? The antecedents of rule following in work settings. Acad Manage J 48(6):1143–1158
Weissmann A (2007) A new approach to corporate criminal liability. Am Crim Law Rev 44:1319–1342

Corporate Liability

► Organizational Liability

Corporate Social Responsibility

► Codes of Conduct

Corruption

Maurizio Lisciandra
Department of Economics, University of Messina, Messina, Italy

Definition

Corruption can be generally intended as an (illicit) exchange between a member of an organization and another subject at the expense of either the organization itself or the rights of others, in which acts of power in contrast with official duty are exchanged for personal advantage. This general definition is usually narrowed to consider corruption as an act of misuse of public power for private profit against the common good. Sensu stricto, bribery is considered the only actual form of corruption. It consists of promising, offering, or giving, as well as soliciting or accepting, a corrupt exchange between some utility (e.g., kickbacks, gratuities, sweeteners) and the actions of individuals (e.g., bureaucrats, politicians, employees in private enterprises) in charge of a legal or public duty. Some habits, such as tipping, gift-giving, and patronage, may not be considered corruption per se, but they are corruption-like situations when they are intentionally conceived to induce someone to act dishonestly. Finally, according to many countries' specific legal standards, certain transfers that are similar to bribes are not considered illicit payments, as in the case with lobbying, campaign contributions, and postretirement offers to politicians or public officials. In these circumstances, the elements of a corrupt exchange are all in place apart from the difficulty of identifying the abuse of the official duty and the distortion of market competition.

Introduction

The phenomenon of corruption was originally a matter of investigation in law and criminology studies. However, corruption is not just a legal or ethical issue but also has important economic implications. Starting with the seminal work by Rose-Ackerman (1978) and becoming a well-established field in the economic literature after the 1980s, the advances in corruption studies have profited from the set of tools and results of the theoretical and empirical economic analysis. The general outcome is that corruption affects both distribution and efficiency and, by its nature, restricts market competition. General discussions and surveys on the advances of corruption studies from an economic perspective can be found in Jain (2001), Tanzi (2002), Svensson (2005), Lambsdorff (2007), Lisciandra (2014), and Aidt (2003, 2016), among many.

Corruption is highly context-dependent, as social norms, culture, institutional setting, and the stage of development affect corruption; however, there are also common patterns in human

behavior that explain what generally causes corruption. Corruption also feeds back on the institutional and economic dynamics to such an extent that consequences and causes are confounded by the complexity of all interactions. These issues are presented in this brief essay, which also provides a conceptual framework of corrupt behavior according to the widely used principal-agent model and a summary of the empirical evidence on the determinants and effects of corruption. However, before moving on to the contributions of the existing literature, it is important to address one important topic which regards the measurement of corruption and is closely intertwined with its definition and the empirical analysis.

Measurement of Corruption

Corruption is a multifaceted phenomenon, hard to delimit and measure. However, many institutions have tried to come up with some measurement that could capture its diverse dimensions within a country, a region, or some economic activity. Currently, data on corruption are abundant and are mainly of two types: objective and subjective.

Objective measures are typically reported crimes, court cases, or convictions of bribery and other corruption-related crimes. Unfortunately, objective data have several shortcomings. They do not necessarily measure the actual level of corruption, but may capture instead the quality of investigative, prosecution, and judiciary departments. Sometimes, it may even be the case that low figures of corruption offenses hide a high pervasiveness of the phenomenon. In addition, this type of data is reliable for comparison only if law enforcement and institutional setup are the same in time and space.

Subjective measures are constructed upon the experience or perception of representative samples of the population or field's experts. In particular, some measures are the results of polls of the general public, while other measures use the information available from expert assessment or a combination of factual data such as laws, regulations, institutional data, and on-site visits. Regarding the weaknesses of subjective measures, especially perception indices, see an extensive analysis in Lambsdorff (2005).

There currently exist numerous cross-national perception measures of corruption. The most widely used measure is the Corruption Perception Index (CPI). This composite index measures perception of corruption in the public sector (i.e., administrative and political corruption) and combines surveys and assessments of corruption from independent institutions devoted to governance and business climate analysis. It is an "index of indices" that was originally launched in 1995 and currently covers 176 countries. Notice, however, that the scores of each country are not comparable over time due to the frequent updates in the methodologies and the inclusion of new data sources.

The International Country Risk Guide Index for corruption is another measure that assesses corruption within the political system in terms of bribery but also secret party funding, excessive patronage, nepotism, and suspicious ties between politics and business. It was originally created in 1980 and derives from the collection of political information that is eventually converted into risk scores. It is issued on a monthly basis and now covers 140 countries. This index is also used to construct the abovementioned CPI.

Another measure of cross-country corruption is the Control of Corruption Index from the Worldwide Governance Indicators that captures the perceptions about petty and grand forms of corruption through a large set of sources such as surveys of firms and households and subjective assessments of several commercial business information providers, as well as many other nongovernmental and governmental organizations. It covers over 200 countries starting from 1996.

There are further subjective cross-country indicators that capture the multidimensional phenomenon of corruption in many of its forms. For a more comprehensive overview of measurements and indicators, see Kaufmann et al. (2007) and Malito (2014). In general, all perception measures appear highly correlated, while, as observed by Mocan (2008), experience-based measures may not be correlated with perception measures due to common perception biases or interviewer biases.

The distinction between objective and subjective measures has important implications in the empirical investigation. Due to the extreme difficulty of finding objective measures that can be comparable across countries (see Escresa and Picci 2015, for a first attempt), empirical analyses adopt subjective measures for cross-country investigations. By contrast, within-country investigations can use both subjective and objective measures. In addition, compared to cross-country investigations, within-country analyses can considerably reduce the institutional differences existing across countries (e.g., administrative controls, investigative departments, criminal laws) that can eventually explain most of the variability in the subjective corruption measures across countries.

Another approach to measurement is through indirect measures of corruption. This is especially the case of corruption in public procurements. Many signals can hide the presence of corrupt transactions, such as delays of work completion, overrun costs, and bad quality of goods and services. For a large overview on this type of measure, see OECD (2007).

Finally, more recently, corruption has also been measured through laboratory and field experiments or so-called natural experiments. These are controlled situations that severely reduce the problem of the identification of causation which exists with both subjective and objective data. Experimental data are typically more suitable to investigate the psychological determinants of corrupt behavior. For a recent survey on the pros and cons of experimental data on corruption and the relevant results, see Lambsdorff and Schulze (2015).

Microeconomic Theory of Corruption

Starting from the 1960s with the analysis of criminal behavior, and exploring from the 1970s onward corrupt behavior in particular, economic theory has provided an important conceptual framework that is useful to analyze many issues involving corruption and its implications.

According to the Beckerian seminal contribution on crime modeling, criminals behave as rational economic agents with a specific risk attitude and decide to breach the law if the utility from breaching is higher than the utility from compliance. Corruption is a two-sided criminal endeavor that follows a similar paradigm. In particular, corruption deterrence would depend on (i) the probability of detection and punishment and (ii) the severity of punishment. These components are rather complementary, to such an extent that a very low level of one of the two makes deterrence extremely weak. The expected cost of committing corruption also depends on other components such as (iii) the psychological cost to infringe some moral or social norms and (iv) the opportunity cost to perform legal activities.

This analytical contribution can be associated with the theoretical analysis of another strand of literature that is widely used in contract theory, the so-called principal-agent model (see Klitgaard 1988). In fact, modern societies are characterized by structured and hierarchical organizations with several delegatory relationships. This type of relationship arises with the advent of division of labor and market exchange. Opportunistic behavior that gives rise to corruption is embedded in many of these delegatory relationships, in which a delegate (agent) exploits the informative advantage with respect to the delegator (principal). In other words, once delegated, the agent is endowed with discretionary power that can be used at the expense of the principal. A typical incident of corruption is when the principal delegates the agent to assign a reward (e.g., a public procurement) or a punishment (e.g., tax collection) to a third subject in the principal's interest, but the agent exploits the informative advantage to his/her own interest, for instance, by pocketing a bribe.

There are two types of informative advantage: adverse selection and moral hazard. The former describes a situation in which the principal does not know some characteristics of the agents (e.g., level of morality), while the latter refers to a situation in which the principal may not perceive correct information on the actions the agent should perform in the principal's interest. Thus,

the principal faces two problems: how to select the agents who are best endowed with morality (i.e., solution to adverse selection) and how to design contract schemes inducing agents to behave honestly rather than dishonestly (i.e., solution to moral hazard).

A solution to the first problem would require that (i) moral virtues should be distributed non-uniformly and that (ii) hiring the best agents should not be very expensive. Unfortunately, moral virtues are not easily observable, and they may not remain constant over time. Nonetheless, there are some indirect signals, such as personal wealth, willingness to work for free, and personal history. All these signals are rather imperfect, especially the first one, and should be used carefully. Consider the problem of adverse selection of political representatives, which is very common in any democracy and gives rise to serious corrupt episodes. Many signals coming from electoral campaigns turn out to be flawed to such an extent that the electoral body (principal) oftentimes takes the bad signals on the morality levels of the candidates (agents) as good ones. In the long run, the general decadence of moral values in politics implies that only the most dishonest individuals would run for elections, while honest individuals would tend to decline any candidacy because they are not willing to reach compromises or to be confused with the dishonest ones. In sum, the worst ones crowd out the good ones. Of course, this is a context of multiple principals in which principals' objectives can be the most diverse, and sometimes a good percentage of the electoral body is not even interested in the moral virtues of candidates. This would eventually exacerbate the problem of adverse selection.

The principal can use the "carrot and stick" method to face moral hazards. The carrot works by aligning the agent's interests to the principal's ones. This can mainly be done, where possible, by contract schemes such as pay-for-performance. The stick is more flexible and aims to increase the expected disutility from sanctions. This can be achieved by increasing the probability of being detected (e.g., monitoring, contrast of interests between briber and public official), the probability of being sanctioned (e.g., investing in the judiciary, extending the statute of limitation), and the severity of sanctions (e.g., easing firing, confiscation, non-electability).

Increasing the community's moral virtues would be another policy in helping to reduce the occurrence of both types of informative advantage. This would mainly increase the psychological cost to breach the law. This is achieved by reducing deprivation, improving law enforcement, introducing rehabilitative punishment, and education around civic-mindedness. In other words, policies that incentivize producers of moral virtues (e.g., the criminal justice system, the education system, volunteering) and discourage producers of dishonesty (e.g., organized crime) would be beneficial against corruption.

Therefore, policies should aim to increase the expected disutility from corruption. For instance, an efficiency wage-like policy would increase the expected disutility from punishment to public officials. This may also increase the potential bribes solicited by public officials, thereby discouraging bribers. However, any policy comes at a cost, which sometimes can be prohibitive. Other policies modify the environment in which corruption takes place, such as (i) reducing public officials' discretionary power by narrowing terms and conditions of their activities, (ii) introducing forms of competition among officials such that more agencies can provide substitutable benefits, (iii) when possible deregulating public purchasing of goods and services, and (iv) simplifying taxation.

Another strand of economic literature that explored corrupt exchange is game theory (e.g., Macrae 1982; Dabla-Norris 2002). Corruption requires an agreement, at least between two parties, and involves strategic interaction between them since they can each cooperate or defect. Cooperation is more likely if defection can be punished. For instance, the briber can punish the uncooperative public officials by excluding them from subsequent corrupt exchanges, hindering their careers, or even encouraging their firing. Evolutionary game theory finds that if corruption is originally confined to small groups that cooperatively sustain the corrupt exchange over time, then corruption is likely to spread to the rest of that

society. In other words, corruption is a sort of disease that is fed by cooperative behavior.

Causes and Effects of Corruption

Corruption is a phenomenon in which causes and effects are closely interdependent. This makes the causation analysis extremely difficult, because the effects of corruption very often feed back into what we expect to be the sources of corruption to such an extent that it is extremely difficult to ascertain which is the cause and which is the effect. In econometrics, this situation is called simultaneity, in which dependent and explanatory variables are jointly determined. For instance, Paldam (2002) draws attention to the seesaw dynamics in which corruption and economic growth feed on each other. For instance, there is a strong negative correlation between country corruption indices and per-capita GDP or per-capita growth rates, but it is not always possible to determine whether corruption generates poverty or whether it is the poverty which is causing corruption. As a consequence, sometimes empirical evidence on the cause-effect relationships is still not conclusive. Nonetheless, some evidence appears to support the strength of a specific causal association. The following empirical evidence, as well as the results from theoretical investigations, presents the current state of the art of the cause-effect relationships of corruption.

Causes of Corruption

A precise and well-defined survey on the causes of corruption can be found in Aidt (2011). Two main factors are considered responsible for corruption and corrupt behavior: institutional or cultural factors and economic drives. On the one hand, weak political and bureaucratic institutions and poor moral virtues encourage the abuse of discretionary power, and on the other hand, economic conditions foster rent extraction. The former are very long-term factors to modify, while the latter can be altered in a shorter period.

Cultural or institutional traditions significantly affect the level of perceived corruption. For instance, protestant religion is a robust predictor of lower perceived corruption (Treisman 2000). Former British colonies also have significantly lower perceived corruption, which is probably due to the presence of common law legal systems. Overregulation and an excess of bureaucracy are found in cultures endowed by excessive social capital, to such an extent that economic activities eventually rely on relationships and connections. Increasing levels of regulation and bureaucracy are positively associated with perceived corruption and may be part of a deliberate strategy of public officials to increase the willingness to pay of private citizens coming across public administration (Schleifer and Vishny 1999). This is a typical example of reverse causality.

Economic factors feed back on corruption. Economic development increases the delegation process of competences and agencies but also the production of goods and services on behalf of the government. A greater involvement of government in the market economy through higher public spending may increase rents as well as discretionary power about rules and resource allocation. The growth of international trade has provided big corporations with the opportunity to explore new markets and to obtain large orders. However, this expansion has sometimes occurred at the price of large bribes paid to officials and politicians of foreign countries, especially in developing economies (Tanzi 2002).

Corruption may also originate from other causes, such as low press freedom, government involvement in promoting industrial policy, widespread poverty, unfair recruitment and promotion procedures, low wages of public officials, and a high tax burden. Finally, endemic corruption is a cause of further corruption because, where corruption is more prevalent, detection and punishment of corrupt episodes is harder, but also the importance of social stigma and loss of reputation decreases (Andvig and Moene 1990; Soares 2004). This is why the level of corruption appears path-dependent.

Effects of Corruption

The main research question on the effects of corruption is whether corruption turns out to be sand

or grease in the complex economic mechanisms of modern economies. This is a controversial question, which has so far produced no definitive answer.

In principle, corruption may enable individuals to avoid bureaucratic delays in order to smooth and speed transactions. According to this line of thinking, corruption is seen as a second-best solution vis-à-vis the slow and inefficient bureaucracy. Especially in the early stages of economic development, inefficient bureaucracy hinders economic growth that could be smoothed by corrupt practices. In the same way, corruption may introduce competition for scarce governmental resources. For example, corrupt practices may minimize the waiting costs for those who place more value on time or could also facilitate firms' entry into highly regulated economies. However, it must be noted that all the grease arguments have undergone several criticisms and have been considered flawed both conceptually and empirically (e.g., see Kaufmann 1997).

The sand argument seems to be corroborated by more robust evidence. As noted in many studies, corruption acts as an uncertainty- and cost-increasing factor. In a general perspective, the main negative consequence of corruption is the impediment to economic growth, especially for those countries with an unreliable institutional framework. Corruption has a negative impact on economic growth through several channels. In particular, it negatively affects both private and public investments. For instance, it discourages investments by potential innovators because established firms are favored in exchange for bribes. Similarly, corruption undermines foreign direct investments, because it acts as a sort of additional tax on business. It weakens the growth-enhancing effects of public investments because corrupt politicians and public officials may tend to divert public resources toward highly profitable investments such as large-scale and high-cost construction projects, which are subject to low rates of return, rather than small-scale and decentralized projects. Corruption also hampers economic growth by distorting individual incentives. It hinders private investments in education by diverting individuals' efforts toward rent-seeking activities and induces bureaucrats to inflate regulations and slow down bureaucratic processes in order to receive bribes from their public administration customers. As seen above, the latter consequence is also a cause of further corruption.

Corruption has been especially investigated for many other undesirable distortions that also feed back on economic growth. The list of negative consequences is very large and cannot be exhausted in this brief overview. For extensive surveys on this topic, see, among many, Lambsdorff (1999) and Pellegrini (2011). Nonetheless, it is worth mentioning additional negative effects, such as the growth of the informal economy, the contraction in international trade, the deterioration of market competition, the increment of income inequality and poverty, and not least the increase in pollution and environmental degradation as well as the weakening of environment policies.

Conclusion

The economic analysis can provide an important analytical framework to explain corrupt behavior and explore its implications. On the one hand, the theoretical approach of the principal-agent model allows us to understand the complex informative set of all subjects involved, their incentives, and the consequent effectiveness of the relevant policies to combat corruption. On the other hand, although measuring corruption is very complicated, an increasingly large number of measures of corruption are available to the public at large as well as to scholars. As a result, econometric analysis has made available an extensive empirical evidence on the cause-effect relationship of corruption, which still appears intricate to disentangle, although a few relevant empirical regularities have been revealed. Of course, this brief entry could only illustrate some of the achievements of such a vast and multifaceted phenomenon as is corruption, which is still far from being fully explored and understood in all its forms and consequences.

Cross-References

- ▶ Administrative Corruption
- ▶ Political Corruption
- ▶ Whistle-Blower Policy

References

Aidt TS (2003) Economic analysis of corruption: a survey. Econ J 113(491):F632–F652

Aidt TS (2011) The causes of corruption. CESifo DICE Rep 9(2):15–19

Aidt TS (2016) Rent seeking and the economics of corruption. Constit Polit Econ 27(2):142–157

Andvig JC, Moene KO (1990) How corruption may corrupt. J Econ Behav Organ 13(1):63–76

Dabla-Norris E (2002) A game theoretical analysis of corruption in bureaucracies. In: Abed GT, Gupta S (eds) Governance, corruption, economic performance. IMF, Washington, DC, pp 111–134

Escresa L, Picci L (2017) A New Cross-National Measure of Corruption. The World Bank Economic Review (2017) 31 (1): 196–219

Jain AK (2001) Corruption: a review. J Econ Surv 15(1):71–121

Kaufmann D (1997) Corruption: the facts. Foreign Policy 107(Summer):114–131

Kaufmann D, Kraay A, Mastruzzi M (2007) Measuring corruption: myths and realities. Africa region findings & good practice infobriefs, no. 273, World Bank, Washington, DC

Klitgaard R (1988) Controlling corruption. University of California Press, Berkeley

Lambsdorff JG (1999) Corruption in empirical research: a review. Transparency International, Berlin

Lambsdorff JG (2005) How corruption Affects Economic Development, Corporate Governance und Korruption. In: Aufderheide D, Dabrowski M (eds) Wirtschaftsethische und moralökonomische Perspektiven der Bestechung und ihrer Bekämpfung. Duncker & Humblot, Berlin, pp 11–34

Lambsdorff JG (2007) The institutional economics of corruption and reform: theory, evidence and policy. Cambridge University Press, Cambridge

Lambsdorff J G, Schulze G (2015) What can we know about corruption?. Jahrbücher für Nationalökonomie u. Statistik 235/2

Lisciandra M (2014) A review of the causes and effects of corruption in the economic analysis. In: Caneppele S, Calderoni F (eds) Organized crime, corruption, and crime prevention. Springer, New York, pp 187–195

Macrae J (1982) Underdevelopment and the economics of corruption: a game theory approach. World Dev 10(8):677–687

Malito D (2014) Measuring corruption indicators and indices. EUI working papers, no. 13, Robert Schuman Centre for Advanced Studies, Florence

Mocan N (2008) What determines corruption? international evidence from microdata. Econ Inq 46(4):493–510

OECD (2007) Bribery in public procurement. Methods, actors and counter-measures. OECD Publications, Paris

Paldam M (2002) The cross-country pattern of corruption: economics, culture and the seesaw. Eur J Polit Econ 18(2):215–240

Pellegrini L (2011) Corruption, development and the environment. Springer, Dordrecht

Rose-Ackerman S (1978) Corruption: a study in political economy. Academic, New York

Schleifer A, Vishny R (1999) The grabbing hand: government pathologies and their cures. Harvard University Press, Cambridge, MA

Soares RR (2004) Crime reporting as a measure of institutional development. Econ Dev Cult Chang 52(4):851–871

Svensson J (2005) Eight questions about corruption. J Econ Perspect 19(3):19–42

Tanzi V (2002) Corruption around the world: causes, consequences, scope and cures. In: Abed GT, Gupta S (eds) Governance, corruption and economic performance. IMF, Washington, DC, pp 19–58

Treisman D (2000) The causes of corruption: a cross-national study. J Public Econ 76(3):399–457

Further Reading

Polinsky AM, Shavell S (2001) Corruption and optimal law enforcement. J Public Econ 81(1):1–24

Shleifer A, Vishny RW (1993) Corruption. Q J Econ 108(3):599–617

Cost of Crime

David A. Anderson

Department of Economics, Centre College, Danville, KY, USA

Abstract

The cost of crime guides society's stance on crime and informs decisions about crime-prevention efforts. While early studies focused on crime rates and the direct cost of crime, the aggregate burden of crime involves a much broader pool of information. Counts of crimes do not indicate the severity of criminal acts or the burden of expenditures to deter crime. Beyond aggregating expenses commonly associated with unlawful activity, a thorough examination of the cost of crime covers such

repercussions as the opportunity cost of victims' and criminals' time, the fear of being victimized, and the cost of private deterrence.

Definition

The direct and indirect costs of crime and its repercussions.

Introduction

The cost of crime influences society's legal, political, and cultural stance toward crime prevention and is part and parcel to the benefits of compliance with legal codes. In the ideal state of compliance, there would be no need for expenditures on crime prevention, no costly repercussions of criminal acts, and no losses due to fear and distrust of others. We will not reach that ideal state, but with knowledge of the full cost of crime, we also know the benefit of eliminating any more realistic fraction of that cost.

Early crime-cost studies focused on particular types of crime, geographical areas, or direct repercussions of crime. The aggregate burden of crime involves a much broader array of direct and indirect costs. The cost of crime includes the opportunity cost of time lost to criminal activities, incarceration, crime prevention, and recovery after victimization. The threat of crime elicits private expenditures on locks, safety lighting, security fences, alarm systems, antivirus software programs, and armored car services. The threat of noncompliance with regulations causes myriad federal agencies to dedicate resources to enforcement. And the implicit psychological and health costs of crime include fear, agony, and the inability to behave as desired.

The largest direct outlays resulting from crime in the United States include annual expenditures of $119 billion for police protection and $85 billion for correctional facilities (Kyckelhahn 2011). (This and all figures are in 2014 dollars.) Several less-visible costs are also substantial. For example, in a typical year, US citizens spend $170 billion worth of time locking and unlocking doors, the psychic cost of crime-related injuries is $106 billion (Anderson 2011), and computer security issues cost businesses $82 billion (FBI 2006).

Estimation Methods

Counts of crimes such as the FBI's *Uniform Crime Reports* (UCR) offer no weights on particular crimes according to their severity. In a period with relatively few acts of arson, for example, society can be worse off than before if the severity of those acts is disproportionately great. From a societal standpoint, what matters most is the extent of damage inflicted by crime, which the UCR does not indicate. Measures of the *cost* of crime convey the severity of crimes – an act of arson that destroys a shed carries a different weight than an act of arson that destroys a shopping center.

In the estimation of crime's cost, approaches with a broad scope capture several types of cost shifting that can stem from crime-prevention efforts. A dual analysis of public and private costs captures the potential for public expenditures to shift the burden of prevention away from individuals and firms. The inclusion of many types of crime captures shifts from one type of crime to another. And the consideration of a broad geographical area accounts for the possibility of crime prevention in one area shifting criminal activity to another.

A basic approach is to tally purchases associated with crime, such as expenditures on deterrence, property replacement, medical care, and criminal justice. Market-based estimates are necessarily incomplete because many of crime's costs, including losses of time, health, and peace of mind, are not fully revealed by purchases.

With the contingent-valuation method, investigators use surveys to estimate values for nonmarket cost components such as fear and pain. These surveys are vulnerable to bias that can result, for example, from the hypothetical nature of survey questions, the objectives of the interviewer who words the questions, or the self-selection of respondents with strong opinions.

Hedonic methods yield estimates of crime-cost components drawn from crime's effect on prices paid for goods or services. For instance, other things being equal, the difference in home prices in areas with low and high crime rates reflects the burden home buyers feel from the greater prevalence of crime.

Cohen (2010) describes a "bottom-up" approach of piecing together each of crime's cost components. Estimates based on market prices, contingent valuation, and hedonic pricing can all play a role in bottom-up calculations. A more holistic "top-down" approach is based on the public's willingness to pay for reductions in crime as stated in responses to contingent-valuation surveys. Anderson (1999, 2011) essentially conducted bottom-up investigations. Examples of top-down studies include Cohen et al. (2004) and Atkinson et al. (2005). Heaton (2010) surveys methods for estimating the cost of crime.

Elements of the Cost of Crime

Crime-cost elements fall into four general categories:

1. Crime-induced production

   Crime leads to the purchase of goods and services that would be obsolete in the absence of crime. With the threat of crime, resources are allocated to the production of security fences, burglar alarms, safety lighting, protective firearms, and electronic surveillance, among other examples of crime-induced production. The growing enormity of crime's burden warrants larger outlays for police, private security personnel, and government agencies that enforce laws. As more criminals are apprehended, expenditures on the criminal justice system and correctional facilities grow. If there were no crime, the resources absorbed by crime-induced production could be used to create gains rather than to avoid losses – $50 spent on a door lock is $50 that cannot be spent on groceries. The foregone benefits from such alternatives represent a real cost of crime.

2. The opportunity cost of time spent on crime-related activities

   Criminals spend time planning and committing crimes, and many serve time in prison. Crime victims lose work time recovering from physical and emotional harm. Virtually everyone beyond early childhood spends time locking and unlocking doors, securing assets, and looking for lost keys. Time is also spent purchasing and installing locks and other crime-prevention devices and watching out for crime, for example, as members of neighborhood-watch groups.

3. Implicit costs associated with risks to life and health

   The psychic costs of violent crime include the fear of being injured or killed and the agony of being victimized. Although the costs associated with risks to life and health are perhaps the most difficult to ascertain, a vast literature is devoted to their estimation. Some direct expenditures on crime prevention are made to address these costs, but preventive measures are limited in their ability to deter crime, so a substantial burden of risks to life and health remains.

4. Transfers from victims to criminals

   Fraud, robbery, and theft cause a loss to the victim, but to the extent that the victim's loss is the criminal's gain, there is not a net loss to society. For this reason, it is useful for crime-cost reports to break out the component of the cost that is purely a transfer. That way, readers can consider the cost to victims as well as the net cost to society without the transfer component. The effect of theft on production may also be a wash – the use of stolen goods often substitutes for the purchase of legal goods, while it is likely that the victims will replace what they have lost. Thus, the transfer of stolen goods does not necessitate additional production of similar items. On the other hand, if low prices on stolen merchandise entice some people to buy items they would otherwise forego, some of these transfers may necessitate additional production.

Findings

Comparisons of crime-cost estimates over time suggest a growing burden in the United States, partly due to true increases in the cost of crime and partly due to the inclusion of a broadening scope of crime's repercussions. An early study by the President's Commission on Law Enforcement and Administration of Justice (1967) placed crime's cost at $158 billion. This estimate includes the direct cost of crimes against persons and property, expenditures on illegal goods and services, and public expenditures on police, criminal justice, corrections, and some types of private prevention.

US News and World Report (1974) estimated a $428 billion crime burden for the United States. That included some private crime-fighting costs, although no breakdown was given. Collins (1994) updated the *US News and World Report* crime study with a cost estimate of $1.82 trillion. Collins included the value of shoplifted goods, bribes, kickbacks, embezzlement, and other thefts among the costs of crime. Collins also expanded the scope of crime-cost calculations to include pain and suffering and lost wages.

Anderson (1999) estimated a $2.5 trillion annual cost of crime in the United States, including transfers of $897 billion worth of assets from victims to criminals. The cost of lost productivity, crime-related expenses, and diminished quality of life amounted to an estimated $1.6 trillion. Anderson (2011) estimated a $3.3 trillion annual cost of crime in the United States, including $1.6 trillion in transfers, $674 billion worth of crime-induced production, $265 billion in opportunity costs, and $756 billion in implicit costs of life and health. This more recent study reflects the crime-related expenditures of the post-9/11 era of heightened security and adds expenditures on investigation services and locksmiths, for which data were previously unavailable.

Estimates of the cost of crime are now available for many countries. For example, Brand and Price (2000) estimate a $125 billion annual cost of crime in the United Kingdom, Zhang (2008) estimates a $110 billion annual cost of crime in Canada, and Detotto and Vannini (2010) estimate a $53 billion annual cost of crime in Italy. These findings should not be used for international comparisons due to differences in methodology, scope, and data availability.

Conclusion

Counts of criminal acts do not capture the scale of crimes or the expense of crime prevention and victim recovery. A focus on the cost of crime allows investigators to gauge the enormity of crimes, measure the financial burden of public and private prevention efforts, and incorporate both direct and indirect effects of crime. A thorough assessment of the cost of crime includes the value of victim losses, the cost of crime-induced production, the value of time lost due to crime, and the costs associated with risks to life and health. In the United States, as criminals acquire an estimated $1.6 trillion worth of assets from their victims in a typical year, they generate an additional $1.7 trillion worth of lost productivity, crime-related expenses, and diminished quality of life.

References

Anderson DA (1999) The aggregate burden of crime. J Law Econ 42(2):611–642

Anderson DA (2011) The cost of crime. Found Trends Microecon 7(3):209–265

Atkinson G, Healey A, Mourato S (2005) Valuing the costs of violent crime: a stated preference approach. Oxf Econ Pap 57(4):559–585

Brand S, Price R (2000) The economic and social costs of crime, vol 217, Home office research study. Home Office, London

Cohen MA (2010) Valuing crime control benefits using stated preference approaches. In: Roman JK, Dunworth T, Marsh K (eds) Cost-benefit analysis and crime control. Urban Institute Press, Washington, DC

Cohen MA, Rust RT, Steen S, Tidd ST (2004) Willingness-to-pay for crime control programs. Criminology 42:89–110

Collins S (1994) Cost of crime: 674 billion. US News World Rep 17:40

Detotto C, Vannini M (2010) Counting the cost of crime in Italy. Global Crime 11(4):421–435
Federal Bureau of Investigation (2006) 2005 FBI computer crime survey. Houston FBI – Cyber Squad, Houston
Heaton P (2010) Hidden in plain sight: what cost-of-crime research can tell us about investing in police. RAND occasional paper. Document OP-279-ISEC
Kyckelhahn T (2011) Justice expenditures and employment, FY 1982–2007 statistical tables. U.S. Department of Justice. December, NCJ 236218
President's Commission on Law Enforcement and Administration of Justice (1967) Crime and its impact – an assessment. GPO, Washington, DC
U.S. News and World Report (1974) The losing battle against crime in America. 16:30–44
Zhang T (2008) Costs of crime in Canada, 2008. Department of Justice Canada, Ottawa, rr10-05e

Further Reading

Anderson DA (2002) The deterrence hypothesis and picking pockets at the pickpockets hanging. Am Law Econ Rev 4(2):295–313
Cohen MA (2005) The costs of crime and justice. Routledge, New York

Cost–Benefit Analysis

Jacopo Torriti[1] and Eka Ikpe[2]
[1]University of Reading, Reading Berkshire, UK
[2]King's College London, London, UK

Abstract

Over the past 30 years, cost–benefit analysis (CBA) has been applied to various areas of public policies and projects. The aim of this essay is to describe the origins of CBA, classify typologies of costs and benefits, define efficiency under CBA and discuss issues associated with the use of a microeconomic tool in macroeconomic contexts.

Definition

Cost–benefit analysis (CBA) is an economic technique applied to public decision-making that attempts to quantify and compare the economic advantages (benefits) and disadvantages (costs) associated with a particular project or policy for society as a whole.

Introduction

Over the past 30 years, cost–benefit analysis (CBA) has been applied to various areas of public policies and projects. Even research on CBA varies significantly and can be classified into two wide areas of work. On the one hand there are studies which have attempted to define the technical-economic reasons underpinning CBA. On the other hand there are studies which carried out empirical evaluations over the performance of samples of CBA.

From a theoretical point of view, CBA has been seen as a tool to increase the quality of regulation and public policy through welfare economics principles and Pareto efficiency. CBA in theory allows for the improvement of social and environmental conditions based on empirical evidence (Sunstein 2002; Koopmans et al. 1964) while improving market competitiveness (Viscusi et al. 1987).

Empirical studies in the area of law and economics have focused on the choice of discount rate, (Dasgupta 2008; Gollier 2002; Lind 1995; Viscusi 2007), the integration of distributional principles (Adler and Posner 1999), the choice of datasets (Morrall 1986; Hahn and Litan 2005), and the performance of different methodologies for monetizing benefits and costs in cases where a market value does not exist (Sunstein 2004; Viscusi 1988). The latter point is of particular interest given the distance between theory and practice and deserves further reflection.

This entry describes the origins of CBA, classifies typologies of costs and benefits, defines efficiency under CBA, and discusses issues associated with the use of a microeconomic tool in macroeconomic contexts.

Origins of CBA

Dupuit, a French engineer, and Marshall, a British economist, defined some of the formal concepts that are at the foundation of CBA. The Federal Navigation Act of 1936 required that the US Corps of Engineers should carry out projects for the improvement of the waterway system when

the total benefits of a project exceeded the costs. This was initiated by Congress, which ordered agencies to appraise costs and benefits when assessing projects designed for flood control as part of the New Deal.

In the 1950s economists tried to provide a rigorous, consistent set of methods for measuring benefits and costs and deciding whether a project is worthwhile. This mainly consisted in applying compensation tests and distributional weights. However, such measures were considered by several economists as a failure (Adler and Posner 1999). Notwithstanding opposition, in the USA, CBA was increasingly applied in an expanding domain of policy areas, often following the rationale that alternative policy appraisal tools were less efficient (Pearce and Nash 1981).

Following some experiences in Scandinavian countries and Canada, the US Executive Order 12291 of 1981 institutionalized CBA as a consistent method for the appraisal of Government policies and regulations, hence marking the beginning of the CBA era (Posner 2000).

Costs

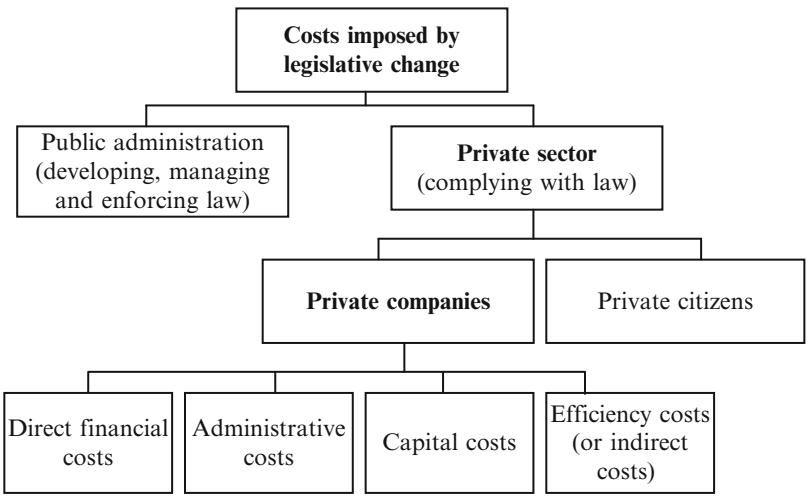
Each type of legislative change imposes various typologies of costs. Private companies, citizens, and public administration can be subject to an increase in costs. The first significant classification is with regard to private and societal costs. The former consist of what a citizen or household has to pay in relation to a legislative change. CBA is often used by public administrations as an instrument to measure only certain components of private costs. This is particularly the case when legislative change is expected to have impacts on individual categories of companies.

Social costs represent what society as a whole has to pay because of legislative change. They typically include negative externalities and exclude transfer costs among groups of citizens (or companies).

Figure 1 outlines the typologies of costs associated with legislative change. Costs for public administration mean management costs as well as enforcement costs, i.e., costs associated with monitoring and inspections to ensure compliance. On the right of Fig. 1, private costs are divided between costs for private citizens and private companies. The latter are broken down in terms of direct financial costs, administrative costs, capital costs, and efficiency costs.

Benefits

CBA practice suggests that benefits are more problematic to quantify and monetize than costs. The following taxonomy of costs outlines some



Cost–Benefit Analysis, Fig. 1 Scheme of costs imposed by legislative change

broad categories of benefits ordered from the highest level of monetization and quantification to the lowest:

- Economic benefits valued in the market
- Noneconomic benefits which are not valued in the market that can be quantified and monetized
- Noneconomic benefits that can be quantified but not monetized
- Noneconomic benefits that cannot be quantified

Economic benefits for which a value is provided in the market are undemanding to monetize. An example could come from adding more wind turbines into the energy market. The market benefits are known because it is known both the physical quantity of energy that the extra turbines would provide (i.e., kWh) and the monetary value of the physical quantity (i.e., €/kWh) (Torriti 2010).

The most controversial category of benefits consists of noneconomic benefits which are not valued in the market that can be quantified and monetized. An example of this is reducing health risks. There is no market value for this and neither for saving lives, but the monetary value of the benefit can be seen as a reduction in the risk of dying or catching a disease. Economists have developed four main methods for monetizing nonmarket values associated with reductions in risk. First, willingness to pay values are based on asking citizens how much they would pay to reduce the likelihood of a specific risk. In practice this is implemented through (i) stated preference surveys, where individuals are asked questions on changes in benefits; (ii) close-ended survey, where respondents are asked whether or not they would be willing to pay a particular amount for reducing risk; and (iii) stochastic payment cards, which offers to respondents a list of prices and associates likelihood matrix describing how likely the respondent would agree to pay the various offered prices. Second, the human capital approach calculates the value of a human life saved, assessing the present value of the worker's earnings over the lifetime. The value is the benefit associated with reducing loss wages. Third, the cost of illness (or medical costs assessment) method consists of an estimate of the costs to the medical system for treatment due to illness. Fourth, willingness to accept values are based on the wage premiums workers accept for risks. When the wage premium is divided by fatality risk, the result is the value of a statistical life saved.

Examples for the noneconomic benefits that can be quantified but not monetized include the number of fish species saved from extinction. CBA is anthropocentric, and impacts on other animal species are rarely taken into account in monetary terms as part of a policy appraisal, unless this refers specifically to ecological conservation and animal species protection. Nonetheless, handbooks on CBA including the British HM Treasury's (2012) Green Book may contain details about parameters to be used for plant species as part of the policy appraisal.

The category of benefits which cannot be quantified and let alone monetized in a CBA is typically very broad, as it comprises several areas of social benefits. An example is the benefit of improving social justice thanks to a new policy or regulatory change. There might be social indicators which address some of this change, but this is hardly reconciled to a monetary value. In the last US Executive Order 13563 on CBA, it is stated that agencies should take account of "human dignity" and "fairness," values, although these are "difficult or impossible to quantify."

Defining Efficiency Under CBA

CBA is the most comprehensive of a family of economic evaluation techniques that seek to monetize the costs and/or benefits of proposals. Following the classifications in the section above, benefits and costs can be broadly defined as anything that increases human well-being (benefits) or anything that decreases human well-being (costs).

The core efficiency principles of CBA lie in welfare economics (i.e., the branch of economic theory which has investigated the nature of the

policy recommendations that the economist is entitled to make) within the domain of allocative efficiency (Baumol et al. 1977; Perman 2003). Allocative efficiency (i.e., allocation of scarce resources that gives maximum social well-being) is defined via the concept of a "Pareto improvement." A Pareto improvement is a reallocation of resources (e.g., a decision to develop) that makes at least one individual better off without making anyone worse off. Pareto efficiency/optimality is achieved when it is impossible to make one individual better off without making at least one other individual worse off.

The problem is that if economists restricted their domain of advice to Pareto improvements, they would not be able to advise on much as most decisions involve a trade-off between making someone better off at the expense of making someone else worse off. What could be required is that the individual who gains must compensate the individual who loses, for all of the latter's loss. If the individual who gains still gains after having paid out the compensation in whatever way (e.g., cash), the move would still be a Pareto improvement. This is not much less restrictive, because actual compensation is rarely paid. As an alternative, economists developed the idea of potential Pareto improvements. The Kaldor compensation test (after Nicholas Kaldor) sanctions a move from one allocation of resources to another, if the winner could compensate the loser and still be better off. In this case, the compensation does not actually have to be paid. Hicks identified a problem with Kaldor's compensation test – namely, that it could sanction a move from one allocation to another, but it could equally sanction a move in the opposite direction, depending on where the problem starts. Instead Hicks suggested that the loser could compensate the winner for forgoing the move, without being worse off than if the change took place. If no compensation takes place, then the reallocation should be sanctioned.

It took a later paper by Scitovsky (1951) to disentangle the problem. Both rules need to be satisfied, such that a reallocation is desirable if on the one hand the winners could compensate the losers and still be better off and, on the other hand, the losers could not compensate the winners for the reallocation not occurring and still be as well off as they would have been if it did occur.

Hence, CBA is based on the Kaldor–Hicks efficiency criterion. The benefits should be enough that those that benefit could in theory compensate those that loose out. It is justifiable for society as a whole to make some worse off if this means a greater gain for others. Under Pareto efficiency, an outcome is more efficient if at least one person is made better off and no one is made worse off. This is a stringent way to determine whether or not an outcome improves economic efficiency. However, some believe that in practice, it is almost impossible to take any social action, such as a change in economic policy, without making at least one person worse off (Buchanan 1959). Using Kaldor–Hicks efficiency, an outcome is more efficient if those that are made better off could in theory compensate those that are made worse off, so that a Pareto-improving outcome results. An allocation is defined as "Pareto efficient" or "Pareto optimal" when no further Pareto improvements can be made.

CBA is a tool for judging efficiency in the case where the public sector supply goods or where the policies executed by the public sectors influence the behavior of private sectors and change the allocation of resources (Stiglitz 2000). The concept of efficiency, though, is normally thought on the premise of the market economy. This is particularly controversial when the decision being contemplated involves some cost or benefit, for which there is no market price or which, because of an externality, is not fully reflected in the market price.

Microeconomic Concepts. . .

As the market economy consists of the spontaneous transaction of goods and services, a characteristic of spontaneous transaction is that there is no individual who loses in the transaction itself. A buyer of goods, who obtains the goods by paying money, is willing to obtain the goods because the amount he or she has to relinquish in exchange for obtaining the goods is in a permissible range. If the amount to be relinquished is

excessive, a person dare not obtain the goods. This assumes that there is an upper limit to the amount of money a buyer is willing to relinquish in exchange for obtaining the goods. In economics, this amount of money is called "willingness to pay" (WTP). No one is willing to buy the goods in the market unless they can be bought at a price lower than WTP. Meanwhile, a seller of the goods, who receives money in exchange for relinquishing the goods, relinquishes them because the amount he or she can get is large enough. A seller will not be willing to relinquish the goods if the amount of money is too small. This assumes that there is a minimum amount of money needed to make a seller willing to relinquish the goods. In economics, this monetary amount is called "willingness to accept" (WTA). No one is willing to sell goods unless they can be sold at a price greater than WTA. In the case where a buyer is the consumer of the goods, WTP for the buyer will depend greatly on the buyer's subjective appraisal of the goods, in other words, the utility the goods bring. WTA is normally equal to the costs of supplying the goods. The agreement of selling and buying in the market means that a buyer bought at a price lower than WTP and a seller sold at a price greater than WTA. A buyer who bought at a price lower than WTP will have the better subjective utility, and a seller who sold at a price greater than WTA should make a profit from the sale that exceeds the costs. That is to say, market transactions do not fail to bring gain to the parties in a transaction. In a case where everyone makes gain or a person makes gain with no one suffering a loss as a result of change, the change is defined as a "Pareto improvement." The market transaction is defined as efficient in that it brings about the Pareto improvement.

. . .Applied to Macroeconomics: The Case of Development Projects

CBA is a method by which this concept of efficiency can be applied to publicly supplied goods as well. Provided that the publicly supplied goods bring utility to people, these people are associated with WTP values for the publicly supplied goods. The WTP of these people represents the benefit of supplying such goods, whereas WTA is the cost for supplying the goods. Difficulty exists, however, in transferring the efficiency concept for market goods to publicly supplied goods. Mishan (1980), for instance, states that the efficiency concept in the meaning of a Pareto improvement cannot be applied because the goods targeted by CBA are supplied goods of a public nature.

Public goods are publicly supplied because they cannot be adequately supplied through the market. It is the reason why they are defined as "public goods." When compared with public goods, private goods have the following characteristics. First, no more than one person can use the private goods at the same time because the process of using the goods is a private process. Second, the private goods cannot be used unless a reward is paid. Third, the users can ordinarily decide whether they use the private goods or not and the degree to which they use them. Public goods do not encompass these characteristics at all (or only to some degree). Among these characteristics, there are two relevant points involving the efficiency concept in CBA: that people can use public goods without paying the reward and that they have no choice as to use or nonuse, or as to the degree of usage. Two critical points in terms of difference between private and social CBA are that (i) public suppliers may be concerned with a much broader range of consequences than firms and (ii) public suppliers may not always use market prices in evaluating projects either because the market prices may not exist or because market prices may not represent true marginal social benefits/costs.

The efficiency criteria on which CBAs for public goods are based are relatively different from the WTP–WTA equation explained above because they are supposed to take into account distribution and equity issues. When efficiency conflicts with other values, it is actually impossible to create economic welfare criteria that integrate all values. Mishan (1982) insisted on the idea that only ethical consensus in society could justify the use of efficiency criteria.

An example of this discussion on the impractical use of CBA in the context of the

macroeconomics of public goods comes from development projects. These create both winners and losers. In most of the cases, losers already belong to the poorest and more marginalized members of society. While development projects can bring enormous benefits to society, their costs to the poorest have effects on their health and even their lives (Kanbur 2002). The use of CBA by international organizations for development projects is widespread and often criticized for not considering areas like basic needs approaches, shadow prices, social discount rates, and macroeconomic shocks to public goods (Brent 1998; Devarajan et al. 1997; Kirkpatrick and Weiss 1996). Examples include how nonmarket values associated with water use and human displacement are considered in CBAs for large-scale hydropower projects in developing countries (Mirumachi and Torriti 2012). Some of these challenges are the result of the paucity of data that thereby requires substantial reliance on approximations and wider resource constraints on assessors (ECA SRO-SA 2012).

Economists tried to introduce sensitive weights to compensate the economic estimates of the project with social, health, and environmental factors. CBA can still be employed in order to identify how losers from, e.g., displacement will be economically affected by the change. The method of aggregation, for example, gives a much larger weight to the gains and losses of the poor than those of the rich. Egalitarianism is introduced through the nature of the utility function. In other words, using the method of aggregation, a dollar's loss or gain means more to a poor person than a rich one. The method of aggregation represents a first move in the direction of the compensation principle (Robbins 1932). Countered (Harrod 1938) and improved (Kaldor 1939) by other economists, the idea of the compensation principle is that a policy change could be Pareto improving if it were accompanied by appropriate lump-sum transfers made by tax winners in order to compensate losers. In practice, the complex combination of using a microeconomic technique in a macroeconomic context means that the systematic use of such weights in project appraisal or CBA is rare.

References

Adler M, Posner E (1999) Rethinking cost-benefit analysis. Yale Law J 109:165–247

Baumol WJ, Bailey EE, Willig RD (1977) Weak invisible hand theorems on the sustainability of multiproduct natural monopoly. Am Econ Rev 67:350–365

Brent RJ (1998) Cost-benefit analysis for developing countries. Edward Elgar, Cheltenham/Northampton, MA

Buchanan JM (1959) Positive economics, welfare economics, and political economy. J Law Econ 2:124

Dasgupta P (2008) Discounting climate change. J Risk Uncertain 37:141–169

Devarajan S, Squire L, Suthiwart-Narueput S (1997) Beyond rate of return: reorienting project appraisal. World Bank Res Obs 12(1):35–46

Economic Commission for Africa Sub- Regional Office for Southern Africa (ECA SRO-SA) (2012) Cost-benefit analysis for regional infrastructure in water and power sectors in Southern Africa. United Nations Economic Commission for Africa, Addis Ababa

Gollier C (2002) Discounting an uncertain future. J Public Econ 85:149–166

Hahn R, Litan R (2005) Counting regulatory benefits and costs: lessons for the US and Europe. J Int Econ Law 8(2):473–508

Harrod RF (1938) Scope and method of economics. Econ J 48(191):383–412

Kaldor N (1939) Speculation and economic stability. Rev Econ Stud 7(1):1–27

Kanbur R (2002) Economics, social science and development. World Dev 30(3):477–486

Kirkpatrick CH, Weiss J (eds) (1996) Cost-benefit analysis and project appraisal in developing countries. Edward Elgar, Cheltenham/Brookfield

Koopmans TC, Diamond PA, Williamson RE (1964) Stationary utility and time perspective. Econometrica 32:82–100

Lind R (1995) Intergenerational equity, discounting, and the role of cost-benefit analysis in evaluating global climate policy. Energy Policy 23:379–389

Mirumachi N, Torriti J (2012) The use of public participation and economic appraisal for public involvement in large-scale hydropower projects: case study of the Nam Theun 2 hydropower project. Energy Policy 47:125–132

Mishan EJ (1980) How valid are economic evaluations of allocative changes? J Econ Iss 14:143–161

Mishan EJ (1982) The new controversy about the rationale of economic evaluation. J Econ Iss 16:29–47

Morrall J (1986) A Review of the record. Regulation 2:25–34

Pearce DW, Nash CA (1981) The social appraisal of projects: a text in cost-benefit analysis. Macmillan, London

Perman R (ed) (2003) Natural resource and environmental economics. Pearson Education, New York /Harlow

Posner RA (2000) Cost-benefit analysis: definition, justification, and comment on conference papers. J Leg Stud 29(S2):1153–1177

Robbins L (1932) The nature and significance of economic science. Macmillan, London
Scitovsky T (1951) The state of welfare economics. Am Econ Rev 41:303–315
Stiglitz JE (2000) Capital market liberalization, economic growth, and instability. World Dev 28(6):1075–1086
Sunstein C (2002) The cost-benefit state: the future of regulatory protection. American Bar Association, Chicago
Sunstein CR (2004) Lives, life-years, and willingness to pay. Colum L Rev 104, 205
Torriti J (2010) Impact assessment and the liberalisation of the EU energy markets: evidence based policy-making or policy based evidence-making? J Common Mark Stud 48(4):1065–1081
Treasury HM (2012) Accounting for environmental impacts: supplementary green book guidance. HM Treasury, London
Viscusi WK (1988) Irreversible environmental investments with uncertain benefit levels. J Environ Econ Manag 15:147–157
Viscusi WK (2007) Rational discounting for regulatory analysis. Univ Chic Law Rev 74:209–246
Viscusi K, Magat W, Huber J (1987) An investigation of the rationality of consumer valuations of multiple health risks. Rand J Econ 18:465

Counterfeit Money

Elena Quercioli[1] and Lones Smith[2]
[1]Department of Economics, College of Business Administration, Central Michigan University, Mt. Pleasant, MI, USA
[2]Department of Economics, University of Wisconsin, Madison, WI, USA

Abstract

Counterfeit money is the topic of television, movies, and lore but hardly seen by most of us – for only about one in ten thousand notes is found to be counterfeit, annually, in the USA (Judson and Porter 2003). And while the value of globally seized and passed counterfeit American dollars has exceeded $250 million in recent years, it is a multi-billion-dollar *potential* crime. Here, we distill a recent model of counterfeit money as a massive multiplayer game of deception. Bad guys attempt to pass forged notes onto good guys. Good guys expend effort to verify the notes, in order to avoid potential losses. The model sheds light on how the counterfeiting rates, the counterfeit quality, and the cost of attention vary in response to changes in the banknote denomination, the technology, and the severity of the legal penalties for counterfeiters.

Synonyms

Deception games; Passed money; Seized money

Introduction

Fiat currency is paper or coin that acquires value by legal imperative. The longstanding problem of counterfeit money strikes at its very foundation, debasing its value and undermining its use in transactions. Williamson (2002) points out that counterfeiting of private banknotes was common before the Civil War but nowadays is quite rare – in fact, only about one in 10,000 notes are counterfeit.

An important distinction must be drawn between *seized* and *passed* counterfeit notes. Seized notes are confiscated before they enter regular circulation, by the police. Passed notes are those fake notes found at a later stage, once they have entered circulation and exchanged hands among unwitting individuals. Passed notes lead to losses by the public, since they must be handed to the police when discovered, and their origins are invariably lost to history.

In the USA, the Secret Service (USSS) investigates the crime of counterfeiting of the dollar. In so doing, it records all seized and passed notes and reports them, annually, to the Congress (US Department of Homeland Security, USSS Annual Report 2012). Since 2002, the European Central Bank (ECB) also reliably records passed and seized notes for the euro, as does the Bank of Canada for the Canadian dollar. Exploring this data, Quercioli and Smith (2014) uncovered some basic facts about counterfeiting, focusing largely on the US dollar.

The data beginning in the 1960s reveal that, initially, the vast majority of all counterfeit notes were seized prior to circulation. But starting

around 1986, this began to change. Currently, the *seized-passed ratio* (the ratio of the number of seized notes and the number of passed notes) has fallen from around 9 to 1/9. But this shift varies by denomination. In particular, the seized-passed ratio increases in the value of the denomination – a pattern that is replicated by the six-denomination Canadian currency.

Next, the *passed rate* (the number of passed notes over the number of circulating genuine notes) rises in the denomination, initially very steeply, and then less so. Notably, for the European currency, the passed rate rises at first and then dramatically plunges at the 500-euro note. Curiously, however, when we look at passed rates for Federal Reserve Banks (referred to as FRBs, hereafter), we notice that they exhibit the opposite pattern: FRBs disproportionately detect previously missed, *low*-denomination notes. Their share of passed notes, instead of rising in the denomination, falls in it – at least until the $100 bill is reached.

As well, Quercioli and Smith (2014) observe some other, interesting empirical patterns: The manufacture of fake notes also varies by denomination. Specifically, the $50 and especially the $100 notes are forged using more sophisticated methods than lesser notes. By contrast, the fraction of fake notes manufactured using "inexpensive" digital means is clearly skewed towards the low denominations (the $5s, $10s, and $20s).

The early literature on counterfeiting was quite small, mainly focused on the money-and-search framework of Kiyotaki and Wright (1989). This general equilibrium theory assumes – in its only margin of explanation – that the price of money (e.g., its purchasing power) adjusts to accommodate supply and demand shocks. Such shocks could, for example, derive from the threat of counterfeit notes circulating in the economy. But that framework has problems accounting for the existence of counterfeiting, let alone explaining its stylized facts. For example, Nosal and Wallace (2007) find no counterfeiting equilibrium when the cost of production of forged notes is high enough. Green and Weber (1996) also find no equilibrium with counterfeiting when agents observe a fixed signal of the quality of the money, before trading with each other.

The most sophisticated paper in this thread is Williamson (2002). This paper assumes that banknotes can be forged at a cost and detected at an exogenous chance. Also, as Williamson (2002) remarks, discounting of private banknotes was common before the Civil War, when counterfeiting of the "Confederate dollar" ran rampant. However, nowadays, counterfeiting is a relatively rare phenomenon. Bills may be declined in payment. For example, "No $100 bills accepted" signs abounded in Ontario after a counterfeiting rash in early 2000 (The Globe and Mail 2002).

The critical, missing variable in all previous literature is the attention that individuals pay to their notes. For instance, Quercioli and Smith (2014) observe that after Canada colorized its notes in 1969–1976, passed rates dropped dramatically across all six denominations of the Canadian dollar. So motivated, they pursue instead a "behavioral" explanation for counterfeiting, that we flesh out here. The first element of the story is a grand *cat and mouse* game of deception played amongst good and bad guys. In essence, illegitimate banknote producers deceive unwitting individuals who, in turn, invest effort to avoid deception. They carefully examine the notes acquired before accepting them. Aside from production and distribution costs, counterfeiters court imprisonment and heavy legal fines when ultimately apprehended by the police. Finally, the police can reduce the passage of bad notes, but do so in a mechanical fashion, aided by the verification efforts of innocent individuals.

In the second element of the story, more careful inspection costs more effort, but reduces the chance of accepting bad money. Inevitably, some good guys are deceived. They then proceed to pass on the bad notes to others – the two parties unwitting partners in a larger, collateral game of deception. The moment a fake note is detected, the holder loses it, by law. This *hot potato* game is one of strategic complements: For the better others check for counterfeits, the more everyone else is motivated to check too, to avoid future losses.

All told, there are two interlinked massively multiplayer games: The cat and mouse game pits

counterfeiters against unwitting individuals and the police, and the hot potato game is the collateral battle fought by unwitting individuals, against each other. The two interlinked games are solved in reverse order. Nash equilibrium is the requisite solution concept – namely, all individuals simultaneously optimize, taking as given others' actions. In the cat and mouse game, Nash equilibrium identifies the verification effort of good guys and the quality level of the forged notes. This yields the verification rate. Next, in the hot potato game, the verification rate identifies the counterfeiting rate.

Building a model of counterfeiting is only part of the task at hand. The next question that needs to be answered is: Does the model shed light on the empirical regularities uncovered? To answer this, we sketch an example of the theory developed in Quercioli and Smith (2014). The full model, data set and discussion of the empirical findings can be found in Quercioli and Smith (2014).

An Economic Model of Counterfeiting

Consider an economy where there are two types of anonymous and unrecognizable agents. Both are risk-neutral. One type of agent is bad, while the other is good. The bad agent can choose to enter the market as a counterfeiter of denomination $\Delta > 0$ notes and can choose the quality $q \geq 0$ of fake notes to produce. For simplicity, we assume his expected production is not a choice variable and simply normalize it to 1. If he enters, his expected profits are zero, after accounting for the anticipated *legal penalty* (or monetary equivalent), Λ, when he is arrested. Counterfeiters also expect to have some *money seized*, and in total, this amounts to $S[\Delta]$, in every period. Bad guys try to distribute everything they produce. To be determined is the *counterfeiting rate* κ, namely, the fraction of all notes that are forged and not authentic.

In this example, good guys simply pass on their notes to other good guys every period, in an unmodeled exchange for goods. If it is fake money, then they lose it. So, note by note, they quickly examine each bill they are handed – expending *effort* $e \geq 0$ in doing so. In turn, this successfully uncovers fake notes with a probability $0 \leq v \leq 1$, the *verification rate*. This probability v reflects the effort e and quality q of counterfeits. We have:

$$e = qv^B$$

where $B \geq 2$.

Thus, the verification rate rises in e and falls in q. The verification rate acts as an implicit price. In equilibrium, knowing the quality, an effort choice is formally equivalent to a verification choice. When handed a note in a transaction, a good guy does not know whether that note is counterfeit or authentic. We thus use an analytic device, allowing the good guy to act *as if* he chooses the verification rate $\hat{v}$ to minimize his losses. Those will occur when he encounters a bad note, he does not recognize it, and he passes it onto a good guy who does. For this, we must introduce a variable κ to capture the *counterfeiting rate.* All such events are independent, and therefore the probability of the joint event is the product of the respective, individual chances of each separate event. His total expected losses in the transaction are:

$$\kappa(1 - \hat{v})v\Delta + q\hat{v}^B$$

including the effort costs of examining the note.

Take the first-order optimality condition in $\hat{v}$, and then impose the equilibrium assumption that all good guys are likewise optimizing, so that $\hat{v} = v$. Then,

$$-\kappa v\Delta + qB\, v^{B-1} = 0$$

Inverting this expression yields the *counterfeiting rate,* κ:

$$\kappa = \frac{qBv^{B-2}}{\Delta}.$$

Observe that while κ, the total stock of counterfeit notes, is not an observable variable, the passed rate is. The theory of Quercioli and Smith (2014) struggles to tease out the behavior of this

important unobserved counterfeiting rate from the passed rate – for they are similar but do not coincide. On a "per transaction" basis, it amounts to $\pi = v\kappa$. So, substituting for κ in π:

$$\pi = v\kappa = \frac{qBv^{B-1}}{\Delta}$$

The passed rate admits a nice economic interpretation. It represents *the ratio of the marginal verification cost and the value of the denomination*. Counterfeiters choose to come into the market and select the quality of fakes they wish to manufacture (the European Central Bank has adopted the catch phrase "feel-look-tilt" in its campaign for the security features of the euro, where tilt refers to the hologram). To be specific, producing a counterfeit note of quality q costs:

$$c(q) = Cq^A$$

where $A > 1$.

We simply assume that anticipated, future *legal costs* are fixed at $\Lambda > 0$. Counterfeiters choose quality to maximize their *profits,* while perfect market competition drives their future profits to zero:

$$(1 - v)\Delta - Cq^A - \Lambda = 0$$

Let us call this the $\overline{\Pi}$-*locus*. In light of the legal penalties, there is a positive minimal counterfeit note that can be profitably counterfeited $\underline{\Delta} = \Lambda > 0$; near such note, the level of verification and quality vanishes. First-order conditions for optimal quality imply what we next call the Q^*-*locus*:

$$ACq^A - \frac{\Delta v}{B} = 0.$$

Together, the optimality and zero-profit equations yield the induced (Nash)-equilibrium verification rate:

$$v = \bar{v}\left(1 - \frac{\Lambda}{\Delta}\right)$$

where $\bar{v} = AB/(C + AB) < 1$.

While the verification rate improves as the note value rises, it is bounded to be below one. Intuitively, some fake notes will always pass into circulation.

In summary, the equilibrium choice variables are:

$$q = \left[(1 - \bar{v})(\Delta - \Lambda)/C^2\right]^{\frac{1}{A}} \text{ and}$$

$$e = qv^B = C^{-2/A}(1 - \bar{v})^{\frac{1}{A}}\bar{v}^B\Delta^{-B}(\Delta - \Lambda)^{B+\frac{1}{A}}$$

We see here that effort and quality ramp up in the note, but effort progresses faster than quality. This raises the verification rate.

Finally, we substitute q into our earlier expression for the counterfeiting rate and find:

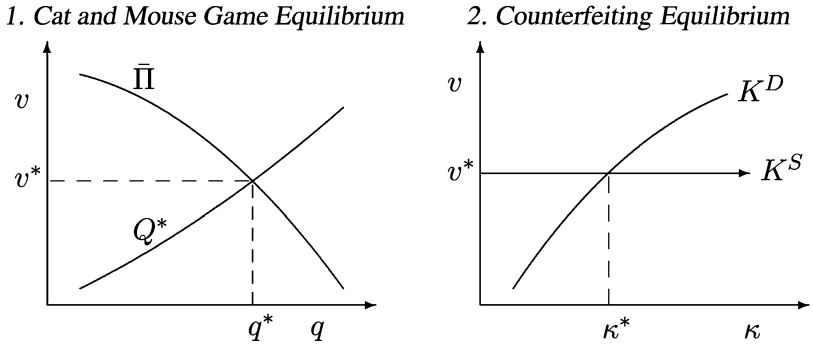
$$\kappa = BC^{-2/A}(1 - \bar{v})\bar{v}^{B-2}\Delta^{1-B+\frac{1}{A}}(\Delta - \Lambda)^{B-2+\frac{1}{A}}$$

Unlike our earlier expression, this formula expresses the counterfeiting rate solely as a function of the exogenously given variables and can be used for predictions.

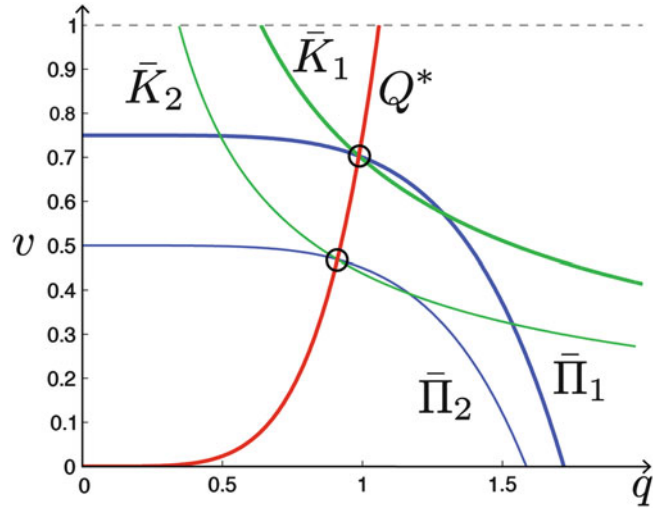
Altogether equilibrium is described by a quadruple of variables (e, q, v, κ) that simultaneously yield equilibrium in the two separate, independent sub-games, linked by the verification and counterfeiting rates. From the cat and mouse equilibrium, drawn on the left hand side of Fig. 1, effort, quality, and thus the verification rate are all determined. Intuitively, one can think of the verification rate as fixing the counterfeiting supply curve K^S – an infinitely elastic curve. (This is so because the homogeneous counterfeiters do not care at all about the counterfeiting rate prevailing in the economy). The "hot-potato" equilibrium then pins down the counterfeiting rate, on the right hand side of Fig. 1, and yields the "derived" demand curve $\kappa = \frac{qBv^{B-2}}{\Delta}$ for counterfeit notes, K^D. It oddly has a positive slope. The reason is that the commodity in question, counterfeit money, is a "bad" and not a "good." So, the verification rate is the price that needs to be paid to discourage illegitimate, counterfeiting activities.

Notice that this equilibrium is *stable*: if the verification rate is too low, below v, counterfeiters will find it profitable to enter the market – as detection is not very effective. The counterfeiting rate surges. On the other hand, if good guys

C



Counterfeit Money, Fig. 1 Equilibrium in the counterfeiting model



Counterfeit Money, Fig. 2 Raising the legal penalty lowers the verification rate and quality

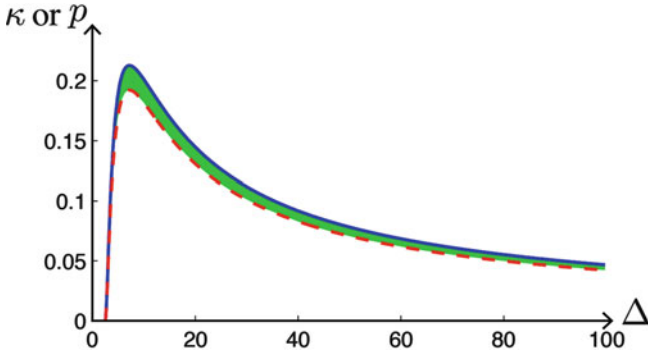
perceive the counterfeiting rate to be below (above) κ, verification ultimately rises (falls). This is so because the good guys realize their errors and readjust their attention level accordingly.

Equilibrium Predictions

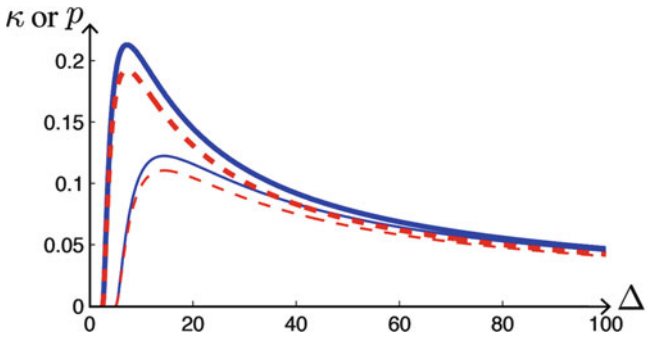
We produce and plot the equilibrium variables. The model exhibits equilibrium feedback. We illustrate one such result of Quercioli and Smith (2014). Suppose the authorities crack down on counterfeiting, thereby rising the (anticipated) legal penalty that counterfeiters court. Then counterfeiters would all exit, unless the cost structure somehow improved. The only variable that can change, here, for a given denomination value, is the effort verifiers expend examining notes. In fact, this must fall so much that – even though counterfeiters produce lower quality notes – the resulting verification rate is lower (see Fig. 2). We find that greater legal penalties strongly "crowd-out" verification effort.

Next, consider the behavior of the counterfeiting and the passed rate. Figure 3, plots them as a function of the denomination. Crucially, this rate is first zero, then rising and eventually falling. This reflects some simple truths. For progressively higher denominations the earlier zero-profit locus $\overline{\Pi}$ and the optimal quality locus Q* shift right, thereby raising quality; but the zero-profit curve moves further to the right and thus verification rises too, see Figure 7, (bottom) in Quercioli and Smith (2014). The inflated verification rate and quality require a greater verification effort too. So, the verification rate, verification effort, and quality all rise in the denomination Δ.

But notice that the counterfeiting rate moves ambiguously. For it is the quotient of the marginal



Counterfeit Money, Fig. 3 Increased legal costs lower verification



Counterfeit Money, Fig. 4 Increased legal penalty lowers the passed rate and counterfeiting rate

verification cost and the denomination value, and both of these rise; however, the first rises proportionately faster at the outset, since it starts at zero. Approaching the lowest note $\Delta \to \underline{\Delta}$ that is still profitable to counterfeit, the counterfeiting rate and passed rate both vanish, since the marginal verification costs of counterfeiting vanish. Loosely, people choose to pay little attention when handed these notes. So, both the counterfeiting rate and passed rate must vanish in this limit too, in order to justify this as an optimal choice. Such inverse reasoning is typical of the strategic analysis in Quercioli and Smith (2014). At large enough notes $\Delta \to \infty$ (a limit that does not obtain for the existing US dollar denominations), this turns around, since the marginal cost of increasing quality rises without bound.

Finally, it is worth observing that the passed rate reaches a maximum at a lower denomination than the counterfeit rate, since their quotient is the verification rate, which is increasing in the note. This is important to keep in mind, in that the most counterfeited note might well exceed the most passed note and certainly is not less.

Now, let us combine our last two lines of thinking, and consider what happens to these whole curves with a rise in the legal penalty Λ. We have seen that *at any given denomination*, this depresses both the verification rate and the quality (see Fig. 2). But our first formulas for the passed rate and the counterfeiting rate are increasing in both the verification rate and the quality. Consequently, both the passed rate and counterfeiting rate curves fall, as seen in Fig. 4.

For a slightly different perspective, observe how Fig. 2 included a third type of curve, namely, a *constant counterfeiting rate locus* $\bar{K}$. This is derived from the optimization of innocent verifiers and consists of all verification rates that would be optimal for a constant counterfeiting rate and the specified quality. Intuitively, this is downward sloping, since a lower verification rate is optimal for a higher quality, holding fixed the counterfeiting rate. This is nicely sandwiched horizontally between $\overline{\Pi}$ and Q^*. Figure 2 shows that the increased legal costs results in a lower quality, and thus the new constant counterfeiting locus $\bar{K}_2$ lies below $\bar{K}_1$.

Quercioli and Smith (2014) also used this graphical apparatus to flesh out predictions for improvements in the counterfeiting technology or in the ease of verification. For instance, cheaper counterfeiting technology lowers the verification rate and raises the counterfeiting rate. In each case, there are feedback effects that undermine but do not overwhelm the intuitive, first-order shifts.

Empirical Evidence via Seized and Passed Money

The total value of *passed money* every "period" is $P[\Delta]$. In fact, the length of the period falls in the velocity of circulation of money. Quercioli and Smith (2014) explore this nuance more carefully, but here it is assumed, for simplicity, to be the average length of time it takes for a note to change hands. Define total counterfeit money produced as $C[\Delta]$. In a steady state, this production must be balanced by the total outflow of seized $S[\Delta]$ and passed $P[\Delta]$ money:

$$C[\Delta] = S[\Delta] + P[\Delta]$$

Also, the amount of passed money in circulation is in balance, and thus its inflow – namely, fakes that escape the attention of the first verifier – must balance its outflow:

$$(1 - v)C[\Delta] = P[\Delta]$$

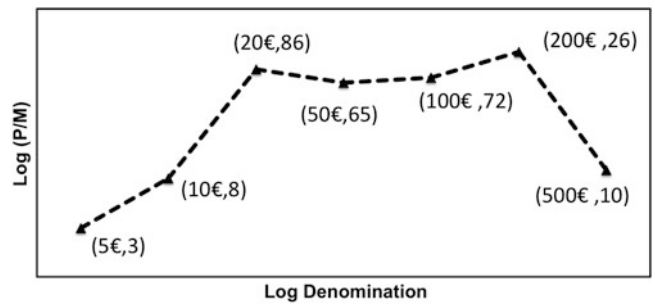
Combining the two expressions reveals the importance of the seized-passed ratio:

$$\frac{1}{1 - v} = 1 + \frac{S[\Delta]}{P[\Delta]}$$

One can now understand that the seized-passed ratio could only have fallen over time if the verification rate had done so too. Quercioli and Smith (2014) identify a digital counterfeiting revolution – the advent of digital color printers, for instance – that strongly suggests a fall in the costs of counterfeiting. That, in turn, would have depressed the verification rate. Additionally, they document the fact that the *counterfeit-passed ratio* $C[\Delta]/P[\Delta]$ rises in the note, but less than proportionately so. For as argued, the verification rate rises in the denomination; on the other hand, higher denominations demand higher quality, in order to deflect increasingly attentive verifiers. All in all, this raises the average costs of counterfeiting. For example, if $100 bills passed half as often as $50 bills, then counterfeiters of $100 bills would lose money. This explains why the counterfeiting revenues, namely, the chance that a note passes times the denomination, must rise in the denomination. In other words, the inverse counterfeit-passed ratio rises in the denomination but less than proportionately, as claimed. We conclude that the model makes sense of the data patterns we observe.

Next, Fig. 5 reports data for the euro counterfeits passed rates from 2002 to 2013 as initially rising and then falling. These data are consistent with the earlier plot depicted in Fig. 3. While the US dollar denominations do not rise high enough to show this, the euro denominations certainly do: notice how the passed rate plummets, starting at the €200 and €500 denominations.

As mentioned in the section "Introduction," a curious inverse piece of evidence for the costly



Counterfeit Money, Fig. 5 The average ratio of passed counterfeit notes to money circulation (in millions) for the euro currency from 2002 to 2013, by the seven euro denominations

attention model is provided by counterfeit money that is missed by verifiers and regular banks. Commercial banks typically pass damaged notes to FRBs. Let the *FRB-ratio* be defined as the counterfeiting rate at FRBs divided by the average counterfeiting rate. Quercioli and Smith (2014) show that this ratio *falls* in the note, except for the $100 note. So, even though $1 is the poorest quality counterfeit, it is disproportionately often found at the FRBs. The reason is that this note secures the lowest verification rate and so it is missed systematically by banks and individuals. Eventually, it surfaces at the FRBs.

One can equally well use the Quercioli and Smith (2014) model to estimate the underlying parameters of the model. Our earlier expression for the seized-passed ratio rearranges to:

$$v = \frac{S[\Delta]}{S[\Delta] + P[\Delta]} .$$

In fact, this is an upper bound on the true verification rate, since the police undoubtedly intercept some counterfeit notes. Calculating the right-hand-side ratio for the US dollar seized and passed data in Quercioli and Smith (2014) reveals that the verification rate upper bounds range from 0.25 (for the $5 note) to 0.31, 0.43, 0.52, and 0.54 (for, respectively, the $10, $20, $50, and $100 notes). This increasing verification rate agrees with the model predictions. It is also interesting to notice the implications of this finding for the current literature on counterfeiting. It unequivocally refutes any assumption that agents observe fixed authenticity signals for notes, as contended, for example, by Williamson (2002). For the verification rate is clearly not fixed, but varies by denomination.

But, what about the so-called street price of a counterfeit note? Logically, this must – at most – equal the average cost of production of a counterfeit note. The conjectured average costs from Quercioli and Smith (2014) by and large agree with the available anecdotal evidence on this. For instance, in one recent case, a Mexican counterfeiting ring sold counterfeit $100 notes at 18% of their face value to distributors, who then resold the counterfeit notes for 25–40% of their face value. The money was transported across the border by women couriers.

In modeling counterfeiting, an important question is whether one can estimate the unobserved counterfeiting rate. As we know, its observable manifestation is the passed rate $p[\Delta] = v[\Delta]\kappa[\Delta]$. Of course, this passed rate is on a "per transaction" basis and not annual. For now, let us assume that the annual velocities of notes are all equal, say, to 10. In this case, the actual passed rates are all 10 $p[\Delta]$. Given the upper bounds discussed above on verification rates (by denominations), we can compute the lower bounds for the counterfeiting rate. Dividing annual passed rates by 10 and then dividing by the verification rates yields the following lower bounds for the counterfeiting rate – for, respectively, the $5, $10, $20, $50, and $100 notes – 0.87, 2.517, 1.733, 1.013, and 1.019 per 100,000. In fact, velocities fall in the denomination – intuitively $1 bills are spent far more often than $100 notes – and thus the observed annualized passed rates should fall relative to the per-transaction rates. That skews the numerical implications of the theory. See Quercioli and Smith (2014) for more details.

The Social Costs of Counterfeiting

We conclude with a final word on the costs imposed on society as a whole by counterfeiting. A passed counterfeit note incurs one – and only one – counterfeiting cost but many verification costs until it is eventually discovered. If it survives inspection with chance $(1 - v)$, then it will be inspected on average $1/v$ times. Let us approximate (thereby overstating) the verification cost by the marginal verification cost times v. Since the counterfeiting rate κ is the ratio of the marginal verification cost and Δ, an upper bound on the total verification expenses for the Δ note in its life is given by $v \times$ (marginal verification cost)/$v = \kappa\Delta$. Now, compare this to the insights of Tullock (1967). He predicts that parties to a transfer, or a theft, of D dollars should be collectively willing to spend up to D to affect the transfer, or the theft. So, the stochastic nature of

counterfeiting holds the actual rent-seeking costs – namely, *the social costs of this crime* – to a tiny fraction of their value.

References

Green E, Weber W (1996) Will the new $100 bill decrease counterfeiting? Fed Reserve Bank Minneap Q Rev 20(3):3–10

Judson R, Porter R (2003) Estimating the worldwide volume of U.S. counterfeit currency: data and extrapolation. Finance and Economics Discussion Paper, Board of Governors of the Federal Reserve System

Kiyotaki N, Wright R (1989) On money as a medium of exchange. J Polit Econ 97(4):927–954

Nosal E, Wallace N (2007) A model of (the threat of) counterfeiting. J Monet Econ 54(4):994–1001

Quercioli E, Smith L (2014) The Economics of Counterfeiting. SSRN ID 1325892:1–24. Conditionally accepted at Econometrica

The Globe and Mail (2002) The unwanted bill: has the C-note lost its currency?. , Saturday, 4 May 2002. www.globeandmail.com

Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. Western Econ J 5(3):222–232

U.S. Department of Homeland Security (2012) United States Secret Service Annual Report

Williamson S (2002) Private money and counterfeiting. Fed Reserve Bank Richmond Econ Q 88(3):37–57

Counterfeiting Models: Mathematical/Economic

Andrea Di Liddo
Department of Economics, University of Foggia, Foggia, Italy

Abstract

Counterfeiting and piracy are illicit activities infringing IPR (intellectual property rights). The market for counterfeit can be divided into two important submarkets. In the primary market, consumers purchase counterfeit products believing they have purchased genuine articles (deceptive counterfeiting). In the secondary market, consumers knowingly buy counterfeit products (nondeceptive counterfeiting).

Counterfeiting has, obviously, consequences on genuine producers and consumers; nevertheless, it can have general socioeconomic effects.

There is a considerable body of theoretical and empirical literature on the mechanisms of counterfeit trade and on the economic and social effects of counterfeiting. A number of the methodological papers are undertaken within the framework of operations research and game theory.

Synonyms

Fake; Forgery; Imitation

Introduction

Counterfeiting and piracy are illicit activities infringing IPR (intellectual property rights).

Counterfeit trademark goods and pirated copyright goods are well defined in the Trade-Related Aspects of Intellectual Property Rights (TRIPS) Agreement, signed in Marrakesh, Morocco, on 15 April 1994. Counterfeit trademark goods are goods which cannot be distinguished in their essential aspects from genuine trademark goods and which thereby infringe the rights of the owner of the trademark. Pirated copyright goods are copies made without the consent of the right holder and that constitute an infringement of a copyright or a related right.

In the present context, according to Abalos (1985), the word "counterfeit" does not cover trade practices as gray market sales, parallel sales, diverted sales, passing off, counterfeiting of money, or money substitutes.

The market for counterfeit can be divided into two important submarkets. In the primary market, consumers purchase counterfeit products believing they have purchased genuine articles (deceptive counterfeiting). The products are often substandard and bring health and safety risks that can be very serious. In the secondary market, consumers knowingly buy counterfeit products (nondeceptive counterfeiting) (OECD 2008).

Typical counterfeited products are luxury goods; pharmaceuticals; and automotive and electrical components. However, almost any product can be counterfeited.

The Economic and Social Consequences of Counterfeiting

Counterfeiting has, obviously, consequences on genuine producers and consumers; nevertheless, it can have general socioeconomic effects.

A number of detailed and enlightening reports and books about the consequences of counterfeiting and piracy have been published by international organizations and scholars. Among them we mention OECD (2008), OECD (2016), and Chacharkar (2013).

The latest estimate from the International Chamber of Commerce (ICC) indicates that the total value of counterfeiting and piracy could reach the staggering level of $2.8 trillion by the end of 2022 (ICC 2017).

The extent of counterfeiting is higher in developing economies also because of the relatively weak enforcement of IP.

Counterfeiters often are linked to organized crime and corrupt government officials in charge of countering their illegal activities.

Usually counterfeit products are cheaper than genuine goods but are of inferior quality. Counterfeiting of medicines, airplane, and auto parts has a detrimental effect on the health and safety of the public. The World Trade Organization estimates that approximately 70 percent of all medicines sold in some African countries are counterfeit.

The owners of the intellectual property may suffer loss of revenues from price pressures, royalties, sales, and brand dilution. Moreover, costs will increase because of legal liability and a higher number of warranty claims.

Counterfeiting harms not only individual consumers and businesses but also the society as a whole. Counterfeiting has effects on tax revenues, government expenditures, and, as a consequence of bribery, the effectiveness of public institutions. Governments also suffer additional costs associated with customs; judicial proceedings; increasing of public awareness; and handling of seized goods.

Counterfeiting also discourages innovation. R&D resources are diverted from creating new technologies for consumers to build up more effective methods to deter counterfeiters.

Destruction of counterfeited items is costly and results in bulky waste; moreover, shoddy fakes can seriously damage the environment (e.g., in the case of chemicals industry).

Counterfeiting affects also employment. Employment shifts from genuine firms to infringing ones, where workers live in less healthy conditions and receive lower wages.

C

Strategies to Fight Counterfeiting

Governments and industries fight counterfeiting on a number of fronts, both independently and, equally importantly, with each other through multilateral institutions and on a bilateral and regional basis. A comprehensive multilateral legal framework has been established within the World Trade Organization (WTO).

A good and enforced system of IPR is necessary to counter counterfeiting and piracy, but it is not sufficient. Measures to limit the free circulation of fakes together with address efforts to fight organized crime are of undoubtedly utility. Advertising campaigns to heighten consumer awareness have often been conducted. The contribution of genuine producers to fight counterfeiting is crucial because of their experience and knowledge and it efficiently complements government action.

Modeling Counterfeiting

There is a considerable body of theoretical and empirical literature on the mechanisms of counterfeit trade and on the economic and social effects of counterfeiting. A number of methodological papers are undertaken within the framework of operations research and game theory. Two excellent literature reviews are provided by Staake et al. (2009) and Cesareo (2016).

According to Staake et al. (2009), contributions can be classified as follows: public perception of counterfeiting and its relevance to management theory and practice; qualitative and quantitative investigations on the consequences of counterfeiting for manufacturers of genuine goods and their supply chain partners; supply-side investigations concerning the production settings, tactics, and motives of illicit actors, and the ways in which their products enter the licit supply chain; demand

side investigations focusing on customer behavior and attitudes in the presence of counterfeit goods; managerial guidelines; and legal and legislative issues concerning different options for IP rights enforcements.

Cesareo (2016), in her book, identify, analyze, and systematize the available research on counterfeiting and piracy published over a 35-year time span (1980–2015).

Among methodological papers, Grossman and Shapiro (1988a, b) can certainly be considered milestones.

Grossman and Shapiro (1988a) focus on the effects of deceptive counterfeiting. They develop an equilibrium model of counterfeit-product trade, incorporating into their analysis both the direct effects of counterfeiting and the induced effects on the behavior of legitimate producers. Their analysis is conducted in a dynamic, two-country model with imperfect quality information and brand-name reputations.

Grossman and Shapiro (1988b) investigate non-deceptive counterfeiting. They develop an equilibrium model of the market for status goods, i.e., those goods for which the mere use or display of a particular branded product confers prestige on their owners. The counterfeiting of a status good, then, deceives not the individual who purchases the product but rather the observer who sees the good being consumed and is mistakenly impressed. The efficacy of two alternative policies is investigated: enforcement and confiscation; tariffs on low-quality goods. Interestingly, the authors prove that it is not true in general that stricter enforcement is welfare-improving.

Supply Side Investigations

Researches dedicated to the supply-side issues of the counterfeit market are of great importance for understanding the way the illicit market operates and how licit brand owners can fight illicit producers.

Qian (2014) provides a theory for brand-protection strategies to reduce counterfeiting under weak IPR. His model incorporates two layers of asymmetric information that counterfeits can incur: counterfeiters fooling consumers and buyers of counterfeits fooling other consumers. One of the theoretical predictions of this study is that counterfeit entry induces incumbent brands to introduce new products. Moreover, better channel management complements a company's own enforcements against counterfeits.

Cho et al. (2015) prove that the effectiveness of anticounterfeiting strategies depends critically on whether a brand-name company faces a non-deceptive or deceptive counterfeiter. Therefore, firms and governments should carefully consider a trade-off among different objectives in implementing an anticounterfeiting strategy.

Dual channel supply-structures have been considered in Zhang and Zhang (2015) as a means of mitigating counterfeiting activities. Adopting the vertical differentiation model, consumers' utility toward the brand name product is described as a function of the price and of the perceived quality. The main finding is that the brand name company should continue to sell, sometimes exclusively, through the general channel despite deceptive counterfeiting.

Yao (2005) considers a market where a monopolist sells a genuine luxury product and counterfeiters illegally copy the product and can enter and exit the market freely. Fines paid by caught counterfeiters are supposed to be pocketed by the monopolist and pegged to the price of the genuine product. Surprisingly, the author shows that it is not always true that the presence of counterfeit products hurts genuine producers. Similar results are found in Di Liddo (2015), but with fines not pegged to the price of the true branded items, and in Buratto et al. (2015) where a dynamic framework is adopted.

Demand Side Investigations

An important field of research addresses awareness, purchase intentions, demographic characteristics, or the attitudes of counterfeit consumers.

Eisend and Schuchert-Güler (2006) review selected studies on the determinants of consumers' intention to purchase counterfeit products and provide a theoretical concept in order to explain the motives when purchasing such goods.

Gentry et al. (2006) investigate product counterfeiting from a consumer search perspective. They prove that factors positively affecting purchase decisions of counterfeits are their low prices, the low investment risks when buying low-cost fake, and, in Western markets, the fun of showing imitation products to friends. In China especially, the potential loss of face when exposed as a counterfeit consumer negatively affects the decision to purchase counterfeit goods.

In order to curb counterfeiting growth, most countries enact codes to punish those counterfeiting firms that are caught. Nevertheless, in Italy and in France purchasers of counterfeit products are also fined and authorities have the right to confiscate their counterfeit items when found. Yao (2015) shows that imposing such penalties reduce demand and hence profit of the legitimate producer under some situations. Under uniform distribution of consumers in product quality estimation, social welfare is reduced. Consequently, counterfeit-purchase penalties employed in some countries are not recommended.

Consumer demand for counterfeit luxury brands is often viewed as unethical, but the demand is also robust and growing. Employing in-depth interviews, Bian et al. (2016) identify the psychological and emotional insights that both drive and result from the consumption of higher involvement counterfeit goods. Moreover, they uncover the coping strategies related to unethical counterfeit consumption.

Legal Issues and Legislative Concerns

A wide range of industries agree that there is severe problem with the protection of IPR throughout the world. The book by Chaudhry and Zimmerman (2013) aims to give the most complete description of various characteristics of the IPR environment in a global context. Authors believe that a holistic understanding of the problem must include consumer complicity to purchase counterfeit products, tactics of the counterfeiters as well as actions by home and host governments, and the role of international organizations and industry alliances.

Cross-References

- ▶ Copyright
- ▶ Imitation
- ▶ Innovation
- ▶ Intellectual Property: Economic Justification
- ▶ Optimization Problems

C

References

Abalos RJ (1985) Commercial trademark counterfeiting in the United States, the third world and beyond: American and international attempts to stem the tide. Boston Coll Third World Law J 5:151–182

Bian X, Kai-Yu W, Smith A, Yannopoulou N (2016) New insights into unethical counterfeit consumption. J Bus Res 69:4249–4258

Buratto A, Grosset L, Zaccour G (2015) Strategic pricing and advertising in the presence of a counterfeiter. IMA J Manag Math 27:397–418

Cesareo L (2016) Counterfeiting and piracy. A comprehensive literature review. Springer, Heidelberg

Chacharkar DY (2013) Brand imitation, counterfeiting and consumers. Centre for Consumer Studies Indian Institute of Public Administration, New Delhi

Chaudhry PE, Zimmerman A (2013) Protecting your intellectual property rights. Springer, New York

Cho SH, Fang X, Tayur S (2015) Combating strategic counterfeiters in licit and illicit supply chains. Manuf Serv Oper Manage 17:273–289

Di Liddo A (2015) Does counterfeiting benefit genuine manufacturer? The role of production costs. Eur J Law Econ 3:1–45

Eisend M, Schuchert-Güler P (2006) Explaining counterfeit purchases: a review and preview. Acad Mark Sci Rev 10:1–25

Gentry JW, Putrevu S, Shultz CJ II (2006) The effects of counterfeiting on consumer search. J Consum Behav 5:245–256

Grossman GM, Shapiro C (1988a) Counterfeit-product trade. Am Econ Rev 78:59–75

Grossman GM, Shapiro C (1988b) Foreign counterfeiting of status goods. Q J Econ 103:79–100

ICC (2017.) https://cdn.iccwbo.org/content/uploads/sites/3/2017/02/ICC-BASCAP-Frontier-report-2016.pdf. Last accessed 30 Mar 2017

OECD (Organization for Economic Cooperation and Development) (2008) The economic impact of counterfeiting and piracy. OECD Publishing, Paris

OECD/EUIPO (2016) Trade in counterfeit and pirated goods: mapping the economic impact. OECD Publishing, Paris

Qian Y (2014) Brand management and strategies against counterfeits. J Econ Manag Strateg 23:317–343

Staake T, Thiesse T, Fleisch E (2009) The emergence of counterfeit trade: a literature review. Eur J Mark 43:320–349

Yao JT (2005) How a luxury monopolist might benefit from a stringent counterfeit monitoring regime. Int J Bus Econ 4:177–192
Yao JT (2015) The impact of counterfeit-purchase penalties on anti-counterfeiting under deceptive counterfeiting. J Econ Bus 80:51–61
Zhang J, Zhang RQ (2015) Supply chain structure in a market with deceptive counterfeits. Eur J Oper Res 240:84–97

Counterterrorism

Marie-Helen Maras
John Jay College of Criminal Justice, City University of New York, New York, NY, USA

Abstract

The European Union has developed several policies and actions to combat terrorism. The EU's counterterrorism strategy is fourfold. It seeks to prevent terrorism, protect people and property against terrorism, pursue terrorists, and respond to terrorism. To achieve this, the EU has implemented measures that seek to disrupt terrorists' operations, prevent radicalization and recruitment of terrorists, secure critical infrastructures and borders, impede terrorists' plans and actions, cut off terrorists' funding and support, improve information sharing, enhance cooperation among agencies, and coordinate responses in the event of a terrorist attack.

Definition

Counterterrorism refers to the measures taken to prevent, deter, pursue, and respond to terrorism. These measures are also designed to protect individuals and property.

Counterterrorism

There is no universally accepted definition of terrorism. However, there are certain elements of terrorism that most governments, politicians, policymakers, practitioners, and academicians agree upon. More specifically, it is generally accepted that those who engage in terrorism use coercive tactics (e.g., threats or the use of violence) "to promote control, fear, and intimidation within the target nation or nations for political, religious, or ideological reasons" (Maras 2012, p. 11). The policies and measures implemented to combat this threat are a form of counterterrorism. Counterterrorism includes a plan of action with specific measures designed to deal with the threat of terrorism.

The EU's counterterrorism strategy is fourfold. Firstly, the EU seeks to prevent terrorism. Here, prevention focuses on reducing the opportunities for engaging in terrorism. Opportunities for terrorism can be reduced by increasing the effort involved, increasing the risks of failure, reducing the rewards of terrorism, and removing temptations, provocations, and excuses for terrorism (Maras 2012, 2013). Most economic analyses of terrorism are based on Becker's (1968) rational choice approach to crime (e.g., Landes 1978; Frey 2004; Enders and Sandler 2006), which holds that an individual will engage in crime if the expected utility of that illicit activity exceeds the gains of engaging in other activities. Pursuant to this approach, terrorists are viewed as rational actors who pursue identifiable goals, decide whether or not to act by examining possible courses of actions in terms of costs and benefits, and assess each action's probability of success or failure. In light of this, countermeasures should make it more difficult for terrorists to achieve their objectives and thus increase the perceived costs of so doing. If the effort required to succeed in a task is raised high enough, it is believed that the terrorists might give up on that task or take longer to execute their operations. Accordingly, an effective counter strategy should focus on ways to frustrate criminals by making it more difficult and risky to commit crime and by reducing its rewards (Frey and Luechinger 2007). Another essential element in preventing terrorism is the implementation of measures targeting the radicalization (i.e., the process whereby individuals adopt extremist views) and recruitment of terrorists (Maras 2012, 2014). There are many different forums within which

radicalization and recruitment take place, for example, prisons and the Internet. The EU counterterrorism prevention strategy seeks to target these environments.

Secondly, the EU aims to protect against terrorism. As part of its counterterrorism strategy, the EU seeks to reduce its vulnerability to terrorist attacks and minimize the impact of an attack should it materialize. Protection involves the collective action to secure critical infrastructure, transportation sectors (e.g., airport, rail, and maritime sectors), and borders. Several new technologies, measures, and programs have been implemented to enhance critical infrastructure protection and border security.

The EU has designated the following sectors as critical infrastructure (CI): communications and information technology, finance, health care, energy, food, water, transport, government facilities, and the production, storage, and transport of dangerous goods (e.g., chemical and nuclear materials) (European Commission 2004). To protect CIs, the European Programme for Critical Infrastructure Protection (EPCIP) was implemented, which identifies critical infrastructure, determines interdependencies of critical infrastructure, analyzes existing vulnerabilities, and provides solutions to protect CIs (European Commission 2007). To facilitate EPCIP, critical infrastructure protection (CIP) expert groups were created, processes to enable the sharing of information on CIP were developed, and the Critical Infrastructure Warning Information Network (CIWIN) was implemented. The CIWIN is a warning network that sends rapid alerts about risks and vulnerabilities to critical infrastructure. Relevant public and private stakeholders are required to share information on critical infrastructure interdependency, the securing of critical infrastructure, vulnerabilities and threats to critical infrastructure, and risk assessments of critical infrastructures.

To protect borders, the passenger name record (PNR) agreement was implemented, which called for the collection and storage of information on all passengers traveling in and out of the EU (European Commission 2010). The type of data obtained and recorded includes contact information, billing information, the travel agency where the ticket was purchased, and any available Advance Passenger Information (API) (e.g., date of birth and nationality) (Directive 2004/82/EC). The Visa Information System (VIS) was also created. This database contains visa application information and biometrics of individuals required to have a visa to enter the Schengen Area (European Commission 2013b). Additionally, the Schengen Information System (SIS) was developed. This system holds information on missing persons and stolen, missing, or lost property (e.g., cars, firearms, and identity documents). It further includes information on individuals who are not authorized to stay or enter into the EU or are suspected of being involved in serious crime (European Commission 2013a). SIS is used by law enforcement, judicial authorities, customs agents, and visa-issuing authorities. SIS II was implemented on April 9, 2013, and includes enhanced functionalities such as the ability to link different alerts (e.g., a stolen vehicle and missing persons alert) and engage in direct queries on the system (European Commission 2013b). Moreover, the European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union (Frontex) was set up to improve the management of the external borders of the EU. Frontex is responsible for the control and surveillance of borders as well as for facilitating the implementation of existing and prospective measures concerning the management of external EU borders.

Thirdly, the EU engages in the pursuit of terrorists across borders. This part of the EU's counterterrorism strategy focuses on impeding terrorists' activities (i.e., their abilities to plan and organize attacks, to obtain weapons, to receive training, and to finance operations), improving information gathering and analysis, enhancing police and judicial cooperation, and combating terrorist financing. For instance, to interdict terrorists, the EU Action Plan for Enhancing the Security of Explosives, the European Bomb Database, and the Early Warning System for Explosives and Chemical, Biological, Radiological and Nuclear (CBRN) material were

developed. Here, the counterterrorism strategy focuses on cutting off terrorists' access to materials that can be used in an attack. The Prüm Treaty facilitated and streamlined the exchange of information and intelligence between law enforcement agencies of the EU Member States. It was created in order to improve EU-wide access to and the exchange of information. Additionally, the Hague Programme removed national borders in data collection, storage, and use, thereby creating an EU-wide right of use of data (Balzacq et al. 2006). To facilitate the transfer of person and evidence between Member States, the European Arrest Warrant (Council Framework Decision 2002/584/JHA) and the European Evidence Warrant (Council Framework Decision 2008/978/JHA) were created and implemented. Furthermore, Directive 2005/60/EC was implemented to prevent the use of the financial system for money laundering and terrorist financing.

Finally, the EU seeks to respond to terrorism. To coordinate responses to acts of terrorism, the EU has implemented several measures. Among the most prominent of which are the Crisis Coordination Committee (CCC) and the ARGUS system (a rapid alert system). The CCC evaluates and monitors the development of crises or emergency situations. Specifically, it identifies issues relevant to the situation and options for decisions and actions. ARGUS links all relevant services of the European Commission during a crisis or an emergency (European Commission 2014). It enables the exchange of real-time information in the event of a crisis and/or foreseeable (or imminent) threat that requires action on the European Community level (European Commission 2005). Europol was established by the Treaty on European Union (Maastricht Treaty) of 1992 and also plays a vital role in responses to terrorism by facilitating information exchange between agencies. This agency is primarily concerned with disrupting criminal and terrorist networks and assisting Member States in the EU in their investigations of criminals and terrorists (Europol 2011).

Overall, in its fourfold strategy, the EU seeks to harmonize existing measures and actions aimed at combating terrorism in order to effectively and efficiently counter this threat. These harmonization attempts, however, need to be cost-effective. Cost-benefit analysis can be used to determine whether counterterrorism resources are best allocated to protect life, human rights, and property. Impact assessments on major policies are used to determine if such an efficient allocation of resources is occurring.

Cross-References

- ▶ Cost–Benefit Analysis
- ▶ Impact Assessment

References

Balzacq T, Bigo D, Carrera S, Guild E (2006) Security and the two-level game: the treaty of Prüm, the EU and the management of threats. CEPS working document no 234

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217

Council Framework Decision 2002/584/JHA of 13 June 2002 on the European arrest warrant and the surrender procedures between Member States

Council Framework Decision 2008/978/JHA of 18 December 2008 on the European evidence warrant for the purpose of obtaining objects, documents and data for use in proceedings in criminal matters

Directive 2004/82/EC on the obligation of carriers to communicate passenger data (2004) OJ L 261

Directive 2005/60/EC of the European Parliament and of the Council of 26 October 2005 on the prevention of the use of the financial system for the purpose of money laundering and terrorist financing

Enders W, Sandler T (2006) The political economy of terrorism. Cambridge University Press, Cambridge

European Commission (2004) Communication from the commission to the council and the European parliament of 20 October 2004 – critical infrastructure protection in the fight against terrorism. COM (2004) 702 final. http://europa.eu/legislation_summaries/justice_freedom_security/fight_against_terrorism/l33259_en.htm

European Commission (2005) Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions – commission provisions on "ARGUS" general rapid alert system. COM (2005) 662 final. http://eur-ex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52005DC0662:EN:HTML

European Commission (2007) Communication from the commission of 12 December 2006 on a European programme for critical infrastructure protection. COM (2006) 786 final (7 June 2007). http://europa.eu/legislation_summaries/justice_freedom_security/fight_against_terrorism/l33260_en.htm

European Commission (2010) On the global approach to transfers of Passenger Name Record (PNR) data to third countries. COM (2010) 492 final (21 Sept 2010). http://europa.eu/legislation_summaries/justice_freedom_security/fight_against_terrorism/jl0043_en.htm
European Commission (2013a) Schengen Information System (SIS). http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/borders-and-visas/schengen-information-system/index_en.htm
European Commission (2013b) Visa Information System (VIS). http://ec.europa.eu/dgs/home-affairs/what-we-do/policies/borders-and-visas/visa-information-system/index_en.htm
European Commission (2014) ARGUS – a general European rapid alert system. http://ec.europa.eu/health/preparedness_response/generic_preparedness/planning/argus_en.htm
Europol (2011) Protecting Europe. https://www.europol.europa.eu/content/page/protecting-europe-21
Frey BS (2004) Dealing with terrorism: stick or carrot? Edward Elgar, Cheltenham
Frey BS, Luechinger S (2007) Decentralization as a response to terror. In: Brück T (ed) The economic analysis of terrorism. Routledge, Abington
Landes WN (1978) An economic study of U.S. aircraft hijacking: 1961–1976. J Law Econ 21:1–31
Maras M-H (2012) Counterterrorism. Jones and Bartlett, Burlington
Maras M-H (ed) (2013) The CRC Press terrorism reader. CRC Press, Boca Raton
Maras M-H (2014) Transnational security. CRC Press, Boca Raton

Court of Justice of the European Union

Roland Vaubel
Universität Mannheim, Mannheim, Germany

Abstract

Quantitative studies are surveyed which indicate that the Court is biased towards political centralization. The econometric evidence shows that this bias is not due to a lack of political independence but to self-selection and vested interest. Various reforms are discussed which would reduce self-selection and vested interest, notably the requirement of judicial experience, delegation of judges from the highest national courts, and the establishment of a separate "Subsidiarity Court."

Definition

The Court of Justice of the European Union is the highest court of the European Union.

Before 2009, the Court of Justice of the European Union used to be called the European Court of Justice or, as in the pre-2009 treaties, simply "the Court of Justice." Its predecessor was the Court of Justice of the European Coal and Steel Community. The Court has its seat in Luxembourg. It is composed of one judge from each member state. The judges are appointed by common accord of the governments for a term of 6 years. Reappointment is possible and frequent. The mean term length has been 9.3 years (Voigt 2003). Every 3 years, some of the judges are replaced. Cases are decided by the full court (27 judges), the Grand Chamber (11 judges), or chambers of three or five judges. Decisions are taken by simple majority except in chambers of three judges or concerning the expulsion of judges. Dissenting opinions are neither written nor published. The Court interprets the European treaties and the secondary legislation of the European Union at the initiative of (i) the other EU institutions; (ii) national governments; (iii) national courts (so-called preliminary rulings); (iv) national parliaments, or chambers thereof, if they claim that the principle of subsidiarity has been infringed by a legislative act; (v) the staff of the EU institutions; and (vi) a natural or legal person if a specific act is addressed, or is of direct and individual concern, to that person or if a regulatory act which is of direct concern does not entail implementing measures. There is also a General Court which, in some cases, serves as a court of first instance. Further details are set out in Art. 19 TEU, Articles 251–281 TFEU, and the Protocol on the Statute of the Court of Justice of the European Union.

The Court has often been called a "motor of integration" – both with respect to market integration and political integration, i.e., centralization. This general impression has been confirmed by several empirical studies. Stein (1981), in a seminal paper, showed that none of the

signatories of the Rome Treaty filed an observation in favor of any of the Court's major centralizing moves, while each of the member states opposed the Court in at least one of them. Jupille (2004, p. 98f.) demonstrated that the Court significantly favors the Commission at the expense of the Council. In 2008, Carruba, Gabel, and Hankla (CGH) analyzed the distribution and effects of such observations in a sample of 3.176 issues over a period of 11 years. Their data shows that the Commission and the net balance of the Council disagreed on 199 (6.3%) of the issues and that the Court sided with the Commission rather than the Council on 137 (68.8%) of these issues (Sweet and Brunell 2010, p. 28). Indeed, comparing the coefficients in CGH's regressions, the probability of a ruling in favor of the plaintiff is *significantly* higher when the Commission is the plaintiff or submits an observation favoring the plaintiff and *significantly* lower when the Commission is the defendant or submits an observation favoring the defendant (Vaubel 2009a, p. 216f.). Using the CGH data, Sweet and Brunell (2010) also show that regardless of the net position of the Council, the Commission's observations significantly affect the Court's decision in the desired direction. By contrast, if the Commission does not file an observation, the balance of the Council does not have a significant effect.

The Court's centralist and centralizing bias has been criticized by many scholars (e.g., Schermers 1974; Stein 1981; Philip 1983; Rasmussen 1986; Weiler 1991, 1999; Bzdera 1992; Burley and Mattli 1993; Garrett 1993; Neill 1995; Bednar et al. 1996, 2001; Garrett et al. 1998; Pitarakis and Tridimas 2003; Sweet 2004; Voigt 2003; Josselin and Marciano 2007; Höreth 2008; Vaubel 2009a, b, c). A court should not propagate a political program. It ought to be an impartial and objective interpreter of the law.

How can the Court's bias towards centralization be explained? The economic approach to law distinguishes between preferences and constraints. Is the Court constrained to decide as it does? The independence of the judges is protected in many ways:

- They are free to decide as they like.
- They enjoy immunity from legal proceedings even after they have left the Court. The immunity of an individual judge may not be waived except by the full Court (Art. 3 Statute).
- "A judge may be deprived of his office or of his right to a pension or other benefits in its stead only if, in the unanimous opinion of the judges and advocates-general of the Court of Justice, he (!) no longer fulfils the requisite conditions or meets the obligations arising from his (!) office" (Art. 6 Statute).
- The number of judges can only be increased by adding new member states or amending the Treaties (Art. 19 TEU).
- The Court's independence cannot be abolished or reduced except by amending the Treaties and the Statute, i.e., by a unanimous decision of the member states including their parliaments and/or electorates.

As for potential conflicts of interest, the Treaties emphasize that judges "shall be chosen from persons whose independence is beyond doubt" (Art. 19 TEU and Art. 253 TFEU). The Statute (Art. 4) adds that "the judges may not hold any political or administrative office" and "may not engage in any occupation, whether gainful or not." Moreover, "before taking up his (!) duties each judge shall, before the Court of Justice sitting in open court, take an oath to perform his (!) duties impartially" (Art. 2 Statute). The fact that the judges are biased in favor of centralization is a violation of their oath.

The sole constraint which may affect their behavior is the provision that in order to be reappointed they require the support of their home government and the common accord of all member governments (Art. 19 TEU and Art. 253 TFEU). Term limitations and reappointments are extremely unusual for the judges of a supreme or constitutional court. Do they constrain the Court? The governments may not know how the individual judges – especially their own appointee – have voted in the chamber. The oath which the judges have to take also obliges them "to preserve the secrecy of the deliberations of the Court." However, if one or more governments

dislike the Court's decisions, they may refuse to reappoint all judges seeking reappointment.

If the reappointment constraint were effective, it would favor the Council rather than the Commission. It would be a barrier to centralization, not a centralizing force. Thus, the Court's centralizing bias cannot be explained by a lack of independence. Moreover, a cross-section analysis of 41 national constitutional or supreme courts (Vaubel 2009a) shows that the share of central government in public expenditure is significantly larger in federal states if the independence of these supreme judges is strongly protected by the constitution. Constitutional judges do not centralize because they are dependent but because they are independent. They are an independent centralizing force, preferring to centralize.

In the same vein, the study shows that public expenditure is significantly more centralized when constitutional amendment is difficult, requiring a supermajority in parliament or several votes. Thus, the courts are more centralist than the constitutional legislator. They centralize more than the people or their parliamentary representatives want.

The only effective constraint on the centralizing tendencies of constitutional or supreme courts is legislative override. However, as Vaubel (2009a) and Sweet and Brunell (2010) have argued in some detail, legislative override is extremely difficult in the European Union. Treaty amendments require unanimity among the member governments and ratification by all parliaments of the member states. Revisions of secondary legislation always require a proposal from the Commission and usually the assent of the European Parliament, both of which share the centralist preferences of the judges. Moreover, the Council has to muster a qualified majority, in some cases even a unanimous vote. So far, there has been only one case of legislative override in the EU – the "Barber Protocol" in the Treaty of Maastricht (Vaubel 2009a; Sweet and Brunell 2010).

Finally, is the Court effectively constrained by the threat of noncompliance as some authors think? Sweet and Brunell (2010) reject this view: "The EU's legal system is organised to deal with non-compliance: member state non-compliance will generate legal actions and non-compliance with any important ECJ ruling will generate new litigation, and new findings of non-compliance" (p. 9).

If the centralist bias of the judges is neither effectively constrained nor due to constraints, it must be attributed to their preferences. Why do they prefer more centralization than the governments, the national parliaments, and the voters do? Two explanations have been discussed (Vaubel 2009a).

The first is the self-selection hypothesis: The experts eligible for appointment to the EU Court are lawyers who believe in centralization rather than subsidiarity. That is why they have been specializing in EU law. They have studied what they cherish, not what they detest.

The second explanation is the vested interest hypothesis: The EU judges, like the Commission and the members of the European Parliament, are interested in centralization at the EU level because it increases their influence and prestige. The larger the powers of the European institutions and the larger therefore the extent of EU legislation, regulation, and administration, the more important and interesting are the cases that the EU judges will be entitled to decide. For example, constitutional courts have to adjudicate interinstitutional disputes at the same level of government. As long as the policy competence belongs to the member states, these disputes are not decided by the EU Court but by the national constitutional courts. But once the competence is transferred to the European level, the European Court is in charge.

These explanations are compatible with each other, and both are compatible with the evidence.

What can be done against centralization by the Court? There are three possible avenues for reform: facilitate legislative override, impose effective constraints on the judges, or alter the preferences of the judges.

Posner and Yoo (2005) argue that independent courts "can be effective only in an institutional setting where external agents such as executive and legislative branches of government . . . correct their errors" (p. 56). Thus, whenever the European Court reinterprets secondary law, the Council may be given the right to reverse the decision without a proposal from the Commission and without the

assent of the European Parliament. Alternatively, if a second chamber consisting of delegates of the national parliaments is added, as the European Constitutional Group (2004) has suggested, this chamber may be given the right of legislative override. Whenever the Court reinterprets the Treaties, the parliaments of the member states or this second chamber may be entitled to correct the judgment.

Could the judges be prevented from passing centralizing judgments in the first place? Again, three proposals come to mind.

First, as Weiler (1999, p. 131) suggests, a qualified majority of the judges may be required to overturn the legislation of the member states.

Second, the Court could be required to publish its voting record or any dissenting opinions. This would enable the governments to better evaluate the performance of their judges. However, judges ought to be independent.

Third, the judges might not be proposed and appointed by the governments of the member states. A survey of 18 modern democracies (Brouard and Hönnige 2010) shows that not a single one has a constitutional court whose judges are exclusively selected by government (s). In four countries, all judges are exclusively chosen by elected parliamentarians, in five countries the majority of the judges is chosen in this way, in six countries all judges have to be accepted by parliament, in one country (Germany) one half of the judges is selected by elected parliamentarians, and in the remaining two a minority of the judges is chosen in this way. In a supranational organization like the European Union, selection by a committee of the national parliaments or by a second chamber of the European Parliament might be appropriate.

Finally, how can the preferences of the judges be changed? Self-selection may be reduced by requiring judicial experience. In the past, only a minority of the lawyers appointed to the Court had previously served in a judicial function in their home country (Kuhn 1993, p. 195). The current president of the Court is a former professor and cabinet minister from Greece. According to the treaties, the judges shall be chosen from persons "who possess the qualifications required for appointment to the highest judicial offices in their respective countries or who are jurisconsults of recognised competence" (Art. 253 TFEU). Moreover, "a panel shall be set up in order to give an opinion on the candidates' suitability to perform the duties of Judge ... The panel shall comprise seven persons chosen from among former members of the Court of Justice and the General Court, members of national supreme courts and lawyers of recognised competence, one of whom shall be proposed by the European Parliament" (Art. 255 TFEU). But this panel has merely a consultative role (Art. 253 TFEU).

The European Constitutional Group (2004) suggests that the EU judges should not only have judicial experience in their home country but also be drawn from its highest court (provided that they have gained judicial experience before being appointed to it). They would be delegated for a term of 8 years after which they would return to their national constitutional court. This would not only minimize self-selection but also improve the competence of the European judges and the integration of EU and national constitutional law. At present, cooperation between the EU Court and the national constitutional courts is also undermined by the preliminary reference procedure (Art. 267 TFEU) which enables the lower courts of the member states to appeal directly to the European Court, bypassing the highest national courts.

The Court's vested interest in centralization is due to the fact that it is responsible for two tasks at the same time: (i) the task of allocating powers between the member states and the European Union and (ii) the task of interpreting EU law within those powers. The solution, therefore, is to separate these tasks and to have two European courts: one court that has no power other than adjudicating cases concerning the division of labor between the member states and the EU – call it the Subsidiarity Court – and one court that decides all other cases. This, too, has been suggested by the European Constitutional Group (2004). The judges of the Subsidiarity Court would be delegated from the highest courts of the member states. This arrangement would

neither invalidate the Court's decisions nor deprive the judges of their independence. It would correct their biased incentives. This is the economic approach.

References

Bednar J, Ferejohn J, Garrett G (1996) Politics of European federalism. Int Rev Law Econ 16:279–294. https://doi.org/10.1016/0144-8188(96)00020-8

Bednar J, Eskridge W, Ferejohn J (2001) A political theory of federalism. In: Ferejohn J, Rackove J, Riley J (eds) Constitutional culture and democratic rule. Cambridge University Press, Cambridge, pp 223–270

Brouard S, Hönnige C (2010) Constitutional courts as veto players. Lessons from Germany, France and the U.S. In: Paper presented at the Midwest Political Science Association conference, Chicago, Apr 2010

Burley A-M, Mattli W (1993) Europe before the court. A political theory of legal integration. Int Organ 47:41–76. https://doi.org/10.1017/S0020818300004707

Bzdera A (1992) The court of justice of the European community and the politics of institutional reform. West Eur Polit 15:122–136. https://doi.org/10.1080/01402389208424925

Carruba CJ, Gabel M, Hankla C (2008) Judicial behavior under political constraints: evidence from the European Court of Justice. Am Polit Sci Rev 102:435–452. https://doi.org/10.1017/S0003055408080350

European Constitutional Group (Bernholz P, Schneider F, Vaubel R, Vibert F) (2004) An alternative constitutional treaty for the European Union. Public Choice 91:451–468

Garrett G (1993) The politics of legal integration in the European Union. Int Organ 49:171–181. https://doi.org/10.1017/S0020818300001612

Garrett G, Kelemen D, Schulz H (1998) The European Court of Justice, national governments and legal integration in the European Union. Int Organ 52:149–176. https://doi.org/10.1162/002081898550581

Höreth M (2008) Die Selbstautorisierung des Agenten: Der Europäische Gerichtshof im Vergleich zum US Supreme Court. Nomos, Baden-Baden

Josselin J-M, Marciano A (2007) How the court made a federation of the EU. Rev Int Organ 2:59–76. https://doi.org/10.1007/s11558-006-9001-y

Jupille J (2004) Procedural politics: issues, interests and institutional choice in the European Union. Cambridge University Press, Cambridge

Kuhn B (1993) Sozialraum Europa: Zentralisierung oder Dezentralisierung der Sozialpolitik? Schulz-Kirchner, Idstein

Neill SP (1995) The European Court of Justice: a case study in judicial activism. European Policy Forum, London

Philip C (1983) La cour de justice des communautées Européennes. Presses Universitaires de France, Paris

Pitarakis J-Y, Tridimas G (2003) Joint dynamics of legal economic integration in the European Union. Eur J Law Econ 16:357–368. https://doi.org/10.1023/A:1025366909016

Posner EA, Yoo JC (2005) Judicial independence in international tribunals. California Law Rev 93:3–74

Rasmussen H (1986) On law and policy in the European Court of Justice. Nijhoff, Dordrecht

Schermers H (1974) The European Court of Justice: promoter of European integration. Am J Comp Law 22:444–464. https://doi.org/10.2307/838965

Stein E (1981) Lawyers, judges and the making of a transnational constitution. Am J Int Law 75:1–27. https://doi.org/10.2307/2201413

Sweet AS (ed) (2004) The judicial construction of Europe. Oxford University Press, Oxford

Sweet AS, Brunell T (2010) How the European Union's legal system works – and does not work: response to Carruba, Gabel and Hankla. Faculty Scholarship series paper, vol 68. Yale Law School

Vaubel R (2009a) Constitutional courts as promoters of political centralisation: lessons for the European Court of Justice. Eur J Law Econ 28(3):203–222

Vaubel R (2009b) The European institutions as an interest group, vol 167, Hobart paper. Institute of Economic Affairs, London

Vaubel R (2009c) The European constitution and interjurisdictional competition. In: Meessen KM (ed) Economic law as an economic good. Sellier European law publishers, Munich, pp 369–381

Voigt S (2003) Iudex calculat: the ECJ's quest for power. Jahrbuch für Neue Politische Ökonomie 22:77–101

Weiler JHH (1991) The transformation of Europe. Yale Law J 100:2403–2483. https://doi.org/10.2307/796898

Weiler JHH (1999) The constitution of Europe. Cambridge University Press, Cambridge

Courts Voluntary Networks

Sylwia Morawska[1], Joanna Kuczewska[2] and Przemysław Banasik[3]

[1]Warsaw School of Economics, Collegium of Business Administration, Warszawa, Poland

[2]Faculty of Economics, University of Gdansk, Sopot, Poland

[3]Faculty of Management and Economics, Gdansk University of Technology, Gdańsk, Poland

Abstract

Although the legal framework for their establishment and operation is identical, courts are marked by diversity. And it is not just about the differences arising from the court's place in the

hierarchy. Courts not only differ in the tangible resources (depending on size) and intangible resources (the knowledge and skills of employees) they possess but also in the organizational culture and the ability to learn (Banasik and Brdulak 2015) and the reputation they have, as well as in the network of contacts (Banasik and Morawska 2016). Courts are embedded in the dense structure of relations with the environment (Czakon 2007), including with other courts. The interorganizational cooperation between courts takes place not only within hierarchical, i.e., regulatory, networks (regulated courts networks) with regard to the tasks imposed by the legislature but also within heterarchical, i.e., voluntary, networks (voluntary courts networks). Courts should strive to harmonize the services they offer, as opposed to companies, where resources together with the core competencies built upon them serve to build a competitive advantage in the market. Courts do not compete for customers on their products or services. Jurisdiction is determined by regulations. What is more, the citizen has the right to the same services in each court. What may help standardization are voluntary courts networks, where courts will exchange good practices, managerial and organizational. Networking can also contribute to the organizational efficiency of courts of general jurisdiction through the rational use of resources and the harmonious interaction of all the elements of the organization.

Networking in Public Management

The rapidly developing science and management practice showed little interest in public sector organizations and the way they were managed. Meanwhile, in recent years, there has been an increased need for theoretical grounds for management in the public sector and in the units forming this sector, as part of the developing a specific discipline of management sciences, i.e., public management. Although management of public organizations becomes, both theoretically and empirically, the object of research investigations and explorations by many scientific fields and disciplines and by practitioners, knowledge in this field is still insufficient. Studies on the behavior of public organizations worldwide in particular relate to public administration (central and local government), schools, or health organizations. In this context, there is a clear cognitive gap as regards the functioning of courts. Courts have so far been considered almost exclusively as a legal entity, with scientific papers concerning them dominated by the law faculties at universities.

Meanwhile, in the judiciary, we can observe phenomena similar to those that take place in other public organizations or in the private sector. This applies inter alia to the networking of courts of general jurisdiction. The network approach has been widely analyzed in the framework of the co-management paradigm. It emerged in the late 1980s and early 1990s of the twentieth century. The issue of network management has been introduced into considerations on management in the public sector by (Heclo 1978; Heclo and Wildavsky 1974). It was taken up by public management theoreticians (Marin and Mayntz 1991; Kooiman 1993; Scharpf 1994; Sorensen and Torfing 2007; March and Olsen 1995; Kickert et al. 1997; Rhodes 1997; Pierre and Peters 2000; Hill and Hupe 2002). However, there is no research on networking the judiciary. The judiciary has the potential to take advantage of the mechanisms of network cooperation. Within its framework, interorganizational cooperation can assume different relationships and interactions (Banasik 2015). Interorganizational cooperation is possible in auxiliary and basic (judicial) activities. Interorganizational cooperation and the possibility of its implementation in the core activities require in-depth research. It is characterized by specificity resulting from the fact that it concerns the administration of justice, where judges are independent. Interorganizational networks in the area of the core activities in civil, criminal, economic cases, etc. in horizontal systems can promote unification of views within homogeneous factual circumstances by obtaining consensus

within different interpretational views and thus build confidence in the judiciary. Creating interorganizational bonds in the auxiliary activities primarily serves the transfer of good practices, managerial and organizational.

Networking in the Judiciary: Research Results

Studies relating to the phenomenon of networking in courts of general jurisdiction in Poland were inspired by the results of pilot implementation of new management methods in courts of general jurisdiction undertaken under the project *PWP Edukacja w dziedzinie zarządzania czasem i kosztami postępowań – case management*. The pilot program was introduced in 60 courts of general jurisdiction. Its task was to change the managerial and organizational practices used in courts. The pilot program was also intended to unleash creativity, both in the managers and in the employees: justices and administrative staff of the courts. It served to initiate cooperation between courts, breaking the hierarchical subordination, within the voluntary hierarchical network. This was an additional unplanned result of the pilot program. The pilot program involved the courts of different hierarchy. Initially, the network consisted of 11 courts. At the end of the project, there were 65 courts cooperating within the network. On completion of the pilot program, courts continued to identify good managerial and organizational practices and within the framework of the resulting network exchanged information and knowledge on the possibilities of implementing management solutions in other courts. The pilot implementation of modern methods of managing the courts was in the nature of a research project of particular importance for the justice administration sector. It became a kind of "experimental field" for testing management methods and techniques. In view of the experiences gained by the public administration when implementing business practices, it was assumed that managerial and organizational "good practices" will be identified in courts, and practice from the business and public administration sector will only support this process. The experiences gained by the public administration were to protect the judiciary against taking hasty decisions with regard to adapting solutions that in practice do not work in the public sector. In cooperation with the managerial staff from the pilot courts, the project developed 24 optimal organizational and technical solutions in the area of managing human resources/finances and information/knowledge ("good practices"), and by way of an experiment, up to 15 of them were implemented in each court taking part in the project on the basis of the implementation path developed. Working groups were set up in the pilot courts – staff groups made up of presidents and directors of the courts (60 presidents and 60 directors). The groups met once a month at the time of the pilot program. Their aim was to exchange information and knowledge between the presidents and directors of the pilot courts, i.e., mutual learning. In addition, for each practice proposed by external experts, thematic working subgroups were established, including the presidents and directors of the courts and employees implementing the practices in the primary pilot program.

During the pilot program, a network was established between the courts. It served the exchange of knowledge between particular entities and the creation of solutions aimed to improve management. It was also a platform for sharing good practices in the substantive area of action. Such a nature of action shares the reasoning represented by (Mandell and Keast 2008, 2009), who believe that the main purpose of the network is to connect its members, facilitating joint activities and learning, and consequently to create new solutions to existing problems. Research into the network formed between the courts shows that it is a highly integrated structure, as evidenced by the density on the level of 0.773, which indicates that on average almost 80% of the actors had the opportunity to cooperate with all the others during work on the project. Similar conclusions can be drawn from the level of the actor's average extent, which indicates that in the course of the project, each of the courts had the opportunity to work

with representatives from approximately 50 other institutions. Also, a high average rate of clustering (a measure of how large a portion of the neighbors of a given actor also neighbor on each other) points to a strong integration of the entities involved in the project. Research indicates that the network created new relationships between the courts, which are in no direct hierarchical dependency relation.

To sum up, the network created as a result of the pilot program:

1. It is voluntary and cooperative (members of the network remain independent and interact only as needed, and the links between them are loose and sporadic; these networks assume a loose form of cooperation, their basic aim being to share knowledge).
2. There is no separate management unit, there is no strategic center, and the organizations take decisions on equal terms.
3. It is at its beginning stage of development, where relations are being built, standards are being established, and courses of action are being set.
4. The networks created as a result of the pilot program are regulated by their participants; thus they are in the nature of heterarchical networks.
5. The network under examination did not develop a strategic center. The network that was created was initiated by 11 presidents of courts of general jurisdiction taking part in the first project.

It follows from the analysis of the results of qualitative research that in the judiciary the development of network collaboration is facing serious constraints. As for the voluntary heterarchical networks, none of the presidents wants to act as a strategic center. Adopting this role by the president may cause a conflict between the president and the Minister of Justice, as in this case the president becomes a visionary and strategist for the community of courts. In this respect therefore, he poaches on the preserve of the Minister of Justice. The centrality and popularity of the strategic center of the network create a potential impact on the individual members of the network and then in turn on the entire network. In addition, acting as a strategic center is a major managerial challenge, because it goes beyond the formal limits of the court. The courts participating in the network remain independent in formal and legal terms. The hierarchical tools used in managing the court are of no use here. Given the restrictions and threats to the strategic center, it seems that the court and its managing president can play the role of an initiating unit, but managing the network as such will require co-decision on the part of the courts co-participating in the network. In heterarchical networks, leadership is developed rather as a result of mutual activities by network participants (Müller-Seitz 2012).

Considerations regarding the strategic center in the newly created voluntary network are particularly important, because on completion of the pilot program the network is slowly dying away. Research shows that the leader of the voluntary courts network determines its existence. There is also need for an outside entity in order for the voluntary courts network to function properly. Its role should be limited to network management and the sharing of knowledge bases, while creating networking initiatives, and should be dependent on partners – the courts. In the case of voluntary courts networks, the present authors are in favor of a mixed solution, i.e., leadership combining the qualities of leadership based on consensus and rotary leadership depending on the unique resources held by the partners. In leadership based on consensus, the decision-making process and the goals are the result of joint activities by the members. Rotary leadership refers to a situation where every network participant has a chance to be the leader for some time. Interorganizational networks in the area of the judiciary are aimed at creating and exchanging knowledge, as well as finding new ways of solving problems.

Cross-References

▶ Cooperative Game and the Law
▶ Heterarchy
▶ Horizontal Effects

References

Banasik P (2015) Organizacja wymiaru sprawiedliwości w strukturze sieci publicznej – możliwe interakcje. E – mentor 2(59):56. http://www.e-entor.edu.pl/artykul/index/numer/59/id/1171

Banasik P, Brdulak J (2015) Organisational culture and change management in courts, based on the examples of the Gdańsk area courts. Int J Contemp Manag 14:33–50

Banasik P, Morawska S (2016) The Courts' public image – the desired direction of change. Int J Court Adm 8(1):2–11

Czakon W (2007) Dynamika więzi międzyorganizacyjnych przedsiębiorstwa. Wydawnictwo Akademii Ekonomicznej im. Karola Adameckiego w Katowicach, Katowice

Heclo H (1978) Issue networks and the executive establishment. In: King A (ed) The new American political system. American Enterprise Institute for Public Policy Research, Washington, p 103

Heclo H, Wildavsky A (1974) The private government of public money. Macmillan, London

Hill M, Hupe P (2002) Implementing public policy: governance in theory and practice. Sage, London

Kickert WJM, Klijn EH, Koppenjan JFM (eds) (1997) Managing complex networks. Sage, London

Kooiman J (ed) (1993) Modern governance: new government – society interactions. Sage, London

Mandell MP, Keast R (2008) Evaluating the effectiveness of interorganizational relations through networks. Developing a framework for revised performance measures. Public Manag Rev 10(6):715–731

Mandell M, Keast RL (2009) A new look at leadership in collaborative networks: process catalysts. In: Raffel J, Leisink P, Middlebrooks A (eds) Public sector leadership: international challenges and perspectives. Edward Elgar, Cheltenham, pp 163–178

March JG, Olsen JP (1995) Democratic governance. The Free Press, New York

Marin B, Mayntz R (eds) (1991) Policy networks: empirical evidence and theoretical considerations. Campus Verlag, Frankfurt-am-Main

Müller-Seitz G (2012) Leadership in Interorganizational networks: a literature review and suggestions for future research. Int J Manag Rev 14:428–443

Pierre J, Peters BG (2000) Governance, politics and the state. St. Martin's Press, New York

Rhodes RAW (1997) Understanding governance: policy networks, governance, reflexivity and accountability. Open University Press, Buckingham

Scharpf FW (1994) Games real actors negative coordination in embedded negotiations. J Theor Politics 6(1):27–53

Sorensen E, Torfing J (eds) (2007) Theories of democratic network governance. Palgrave Macmillan, Basingstoke/New York

Craft Guilds

David Dolejší
Faculty of Social and Economic Studies, Jan Evangelista Purkyne University, Ústí nad Labem, Czech Republic

C

Abstract

Craft guilds were formal professional organizations in medieval and early modern Europe that operated in specific areas of industrial production. The primary objective of these organizations was to protect interests of their members. To achieve this goal, craft guilds engaged in a variety of activities from which obtaining legal monopolies over local production and trade within their crafts was the most important.

Definition

Craft guilds were geographically restricted government-licensed organizations of professionals who were specialized in certain areas of industrial production and who were concerned with securing welfare of their members mainly through possessing exclusive rights over the matters of their professions. These cooperatives were found in various forms across the world including China, Japan, the Middle East, Latin America, and North Africa, but their histories are mostly connected with medieval and early modern Europe.

European Craft Guilds

In premodern Europe, craft guilds were a prevailing force in most industries for more than 500 years. Guilds first appeared in antiquity but their heyday came about later in the late medieval period, and in some economies, they continued until the early modern period. Guilds' end came with the rise of national states as governments passed legislation that abolished them.

Premodern professionals formed guilds in almost all spheres of production and trade representing a large part of population. Nearly all urban areas of Europe, and sometimes rural areas, had industries controlled by these cooperatives. In most cases, there were guilds of manufacture producers. These organizations included traditional professions such as furriers, smiths, cobblers, or tailors but also nontraditional professions such as soap makers, dyers, and saddlers. Aside from these manufacturers, there were also widespread guilds of food producers. These included among others brewers, butchers, and bakers. Finally, there were guilds of servicemen, which included most notably merchants and retailers but also occupations such as barbers, painters, surgeons, and drivers. Besides single-craft guilds, some guilds combined affiliated professions in one organization creating multi-occupational guilds.

Although guilds could emerge in any industry, the number of guilds varied from city to city. Some towns had a relatively low density of guilds around one guild per 1000 inhabitants. Others had a relatively large density of guilds over five guilds per 1000 inhabitants. Also the number of craftsmen in each guild varied from a few men to over a 1000 (Lucassen and Prak 1998; Ogilvie 2014).

Activities of Craft Guilds

In medieval and early modern cities, the absence of traditional kinship networks and modern state institutions exposed the newly rising class of professionals to a variety of uncertainties typical to the premodern era. The response was to establish a new form of a network, one that was based on occupational affiliation rather than family ties, a craft guild. Indeed, despite the great variation between regions and even within the same city, the underlying purpose of the guild was everywhere the same: to secure stable standards of living for its members in the face of new risks brought as a result of market development. Craft guilds engaged in a variety of activities in order to achieve this goal.

Private Activities

What made craft guilds different from other forms of urban cooperatives and social networks was their emphasis on economic relationships. Most importantly, guilds laid down conditions for controlling trade and production within a particular geographic location (Ogilvie 2007, 2014). They held the monopoly over production in particular crafts as well as the monopsony over hiring workforce and purchasing inputs. Guilds' market monopolies were fortified by government privileges. Guilds regularly bargained with sovereign authorities over legal privileges, which guaranteed their members exclusive rights over local markets. These privileges enabled guild members to control economic activities of nonmembers within the city and its close neighborhood. Nonmembers were either excluded from trade and production within the particular location or their economic activities were regulated. In either case, guild members were clearly in a favorable position. By obtaining legal privileges, guilds provided their members with economic rents at the expense of nonmembers and the society in general.

However, guilds' concerns did not stop with obtaining economic rents. Guilds used their legal rights over local markets to reduce the transaction costs associated with premodern production and trade (Gustafsson 1987; Epstein 1998; Epstein and Prak 2008). In these respects, craft guilds were able to enforce contracts within their crafts, maintain reputation of their products and services, and provide skill transfers across time and places through a system of apprenticeship and journeymenship.

Furthermore, local experiences indicate that craft guilds were extensively active in areas of life other than economic (Epstein and Prak 2008; Lucassen et al. 2008; Richardson 2005; Prak et al. 2006). Many guilds engaged in the spiritual salvation of their members. The guilds' religious endeavors included organizing members' funerals, maintaining altars, lighting candles, and managing masses and prayers, all secured from collective resources. Guilds served as insurance networks. They provided resources and collective support to masters and their families in difficult situations, especially in sickness, widowhood, or old age. Guilds and their members participated in local politics. The guilds' influence, although limited, was

exhibited through their members, who regularly held positions in local and sometimes national governments.

Public Activities

Although craft guilds were primarily concerned with the interest of guild members, they were government-licensed organizations. Guilds' existence was directly linked with the privileges they obtained from local or national authorities. But authorities did not recognize guilds for free. In exchange for legal rights, governments demanded services from their recipients in return.

Before the state established strong public institutions, craft guilds provided means for civil administration (Hickson and Thompson 1991; Prak et al. 2006). Collection of fiscal revenues was among the most important functions. Guilds' knowledge of local conditions and expertise in craft fitted this function perfectly. Guild members were required to collect a variety of industrial and commercial taxes as well as provide loans and pay dues to their rulers.

Craft guilds also provided a number of civic tasks that benefited the public at large. Guilds' assemblies functioned as courts of the first instance in matters of their professions. Guilds also provided fire services. City charters demanded that craft guilds acquired their own fire equipment such as buckets, hooks, and water pumps and if necessary to engage in firefighting. Guild members also had to patrol inside city walls and finance military activities of their rulers. In some cases, guild members were even required to take part in the defense of their towns.

The Impact of Craft Guilds

Craft guilds were multifunctional organizations. Therefore, it is difficult to pin down their clear impact on society. The traditional literature argues that craft guilds had a mostly negative impact on the economy because of guilds' monopoly of trade and production. However, the new literature on craft guilds tries to moderate this view by showing that craft guilds also had a positive impact on the economy as they provided their members with social insurance, organized funerals, participated in local governance, patrolled on city streets, or pioneered fire services. Missing either of these aspects of guilds may lead to the wrong conclusions. Including private and public dimensions of guilds' activities into the picture is, therefore, important for understanding the role of these organizations in history.

References

Epstein SR (1998) Craft guilds, apprenticeship, and technological change in preindustrial Europe. J Econ Hist 58(3):684–713

Epstein SR, Prak MR (eds) (2008) Guilds, innovation and the European economy, 1400–1800. Cambridge University Press, Cambridge

Gustafsson B (1987) The rise and economic behaviour of medieval craft guilds an economic-theoretical interpretation. Scand Econ Hist Rev 35(1):1–40

Hickson CR, Thompson EA (1991) A new theory of guilds and European economic development. Explor Econ Hist 28(2):127–168

Lucassen J, Prak MR (1998) Guilds and society in the Dutch Republic. In: Núñez CE (ed) Guilds, economy and society. Universidad de Sevilla, Sevilla, pp 63–78

Lucassen J, de Moor T, van Zanden JL (eds) (2008) The return of the guilds. Vol. 53. International Review of Social History, Suplement 16

Ogilvie S (2007) 'Whatever is, is right'? Economic institutions in pre-industrial Europe. Econ Hist Rev 60(4): 649–684

Ogilvie S (2014) The economics of guilds. J Econ Perspect 28(4):169–192

Prak MR, Lucassen J, Soly H, Lis C (eds) (2006) Craft guilds in the early modern low countries: work, power and representation. Ashgate, Aldershot

Richardson G (2005) Craft guilds and Christianity in late-medieval England: a rational-choice analysis. Ration Soc 17(2):139–189

Creative Commons and Culture

Joëlle Farchy[1] and Pierre-Carl Langlais[2]
[1]Faculty of Economics, Pantheon-Sorbonne University Paris 1, Paris, France
[2]Université Paris IV – Sorbonne, Paris, France

Inspired by a specific philosophy, the "free" movement first emerged in the software and academic research fields and has since enjoyed

considerable economic success. Originally, the development of free software was in keeping with practical considerations. Computer specialist Richard Stallman, the "father" of free software, one day in the late 1970s became enraged at his rebellious desktop printer, because he was unable to access the program that controlled it or get it to respond to his commands. The real story begins in 1983, when he sent out a message announcing the creation of a comprehensive software package compatible with the proprietary system, UNIX. He intended to make it available to anyone who wished to use it. . .for free. This project was called GNU. In 1991, Linus Torvalds, a Finnish student, perfected Linux, the first version of an entire operating system based on GNU. The Free Software Foundation (FSF), a nonprofit association, was established in 1985 by Stallman to ensure the logistical structuring and financing of the "free" project, GNU, which then gave rise to GPL (General Public License). In the software field as in that of scientific publications, digital technology led to the emergence of systems combining intellectual property rights and extended tolerances for certain types of use.

In the field of culture, initiatives are born in a specific intellectual context, which we begin by briefly reviewing, and the revolutionary ambitions of pioneers have evolved with the development of Creative Commons licenses (1). These licenses, which originally were but one form of free license among many, have gradually gained a monopolistic position for numerous reasons, and in their 15 years of existence, their use has greatly increased (2). The purpose of this article is to make an initial assessment of the use of these licenses based on the scattered statistical data that is available. This appraisal provides an overview in terms of the works available (3), the varied practices of creators and consumers of works (4), and different economic models (5).

A Revolutionary Initial Goal

"Free culture" (the term was used by Lessig for the first time in 2004 but has its roots in earlier movements) is the heir to a double movement – a community ideal characteristic of the digital utopias of the late 1960s and the sharing practices promoted by theoreticians of the commons. The convergence between these two movements occurred in several American university centers in the 1990s.

Digital utopias are a first breeding ground for thinking that emphasizes community ideals and self-regulation. Starting in the 1960s, certain American counter-culture movements began using digital technologies. Computers, until then more often associated with a public and private managerial planning, appeared as a vector of emancipation. New Communism, notably, based on the idea of small, self-managed communities, influenced an entire generation of entrepreneurs and intellectuals (Turner 2006) but did not, however, contradict the development of intellectual properties and use restrictions. In contrast, the principle of commons, based on Elinor Ostrom's founding work (1990) on self-managed agricultural communities and initially far from digital technology, stressed the sharing of resources and collaborative governance. The 2009 Nobel Prize in Economy Winner highlighted the sustainability of forms of governance that are neither public nor private, where resource allocation is not based not on the individual attribution of property titles and collaborative management of community-pool resources determines the use conditions of these resources.

In the mid-1990s, Commons and digital technology combined and took shape within a small network of academic institutions. The Berkman Center (Harvard University) in particular brought together a number of actors, ideas, and movements closely linked to the sharing of works and information on the Internet: free software (Richard Stallman of MIT was a frequent visitor there), the decentralized networks theory (Yochai Benckler, a Harvard-educated professor at New York School of Law), a positive definition of the public domain (James Boyle, also a Harvard graduate and professor at Duke Law School), and legal analysis of the laws and uses of cyberspace (Lawrence Lessig). The desire to revolutionize, through practice, the copyright rules to facilitate the conditions of the sharing of works gave rise to a plethora of initiatives. The end of the 1990s

marked the golden age of free licenses in the cultural field; the vast majority of licenses were created between 1997 and 2001 (De Filippi and Ramade 2013).

In 2001, Creative Commons was but one free license among others in an already busy sector. Similar initiatives existed: Wikipedia adopted the GNU's GFDL; the Free Art license responded to national specificities in countries with a copyright tradition. On May 16, 2002, a press release announced the launch of a new organization, Creative Commons, in which American lawyer Lawrence Lessing would play a key role by contributing to the project's reorientation. . .and success. Under his leadership, one of the organization's initial key goals was to unite enough people and circles to establish itself as a cultural and social movement, and not just another free license among dozens of others. The radical opinions voiced by early activists gradually gave way to a more conciliatory approach so as to broader future coalitions.

Lessig (2001) developed a new understanding of culture as a common in the digital age. For the latter, the protection that intellectual property affords creators was not be called into question but rather balanced by user rights. Advocates of free licenses simply wish to see copyright used differently to promote the sharing and reuse of commons, meaning that although owners of copyrights retains the legal monopoly, they can use their power to authorize their use rather than forbid it. Licenses offer creators contractual tools that favor sharing and collective creation in the digital world. Free licenses may be used and shared openly, but only under particular terms and conditions, the contours of which must be determined by the individual rights holder who define the availability of their works and the terms of their use by way of "private ordering." Born under revolutionary auspices, Creative Commons gradually built a balance between the anarchic dissemination of works on the Internet and complete control. Creative Commons licenses are thus a revisited approach to intellectual property as a bundle of rights rather than a revolutionary system – in other words, a complement rather than an alternative to copyright.

The Gradual Emergence of the Creative Commons Monopoly

The success of Creative Commons licenses in the cultural field hinges on three strategic choices that have made them indispensable tools for expanding use rights: the implementation of a flexible, modular design, easy access visual symbols and the aim of adapting to national legal specificities despite their international scope.

A Flexible, Modular Design

In 2001, creators who wanted to share their productions could choose between several licenses ranging from total liberalization (WFTPL) to protection against enclosures (GFDL) to full retention of Moral rights (Free Art license). Most of these licenses were monolithic however, offering but a single legal framework to be taken or left. In contrast, Creative Commons offered a modular structure, leaving users free to choose the option best suited to their specific needs. While the core of the initial 2002 version remains unchanged, many adjustments have taken place over the years. In 2016, four options (allocation, identical sharing, noncommercial, and nonderivative) combined to create six separate licenses, to which the public domain was added (see Box 1).

Box 1 *Creative Commons* Licences in 2016

CC-By (By Yourself): The reuser shall give appropriate credit, provide a link to the license, and indicate if changes were made. This is a default option since 2004.

CC-By-SA (*Share Alike*): The reuser shall give appropriate credit and distribute the contributions under the same license as the original (in case of remix or transform of the original work).

CC-By-NC (*Noncommercial*): The reuser shall give

(continued)

Box 1 *Creative Commons* Licences in 2016 (continued)

appropriate credit and may not use the material for commercial purposes. The extent of "commercial use" remains currently fuzzy.

CC-By-ND (*Nonderivative*): The reuser shall give appropriate credit and may not distribute [any] modified material.

CC-By-NC-SA (*Noncommercial/Share Alike*): The reuser shall give appropriate credit, may not use the work for commercial purpose, and shall publish every derivative work under the same license.

CC-By-NC-ND (*Noncommercial/Nonderivative*): The reuser shall give appropriate credit, may not use the work for commercial purpose, and may not distribute any modified material.

CC0 (Public Domain): The work is dedicated to the public domain by waiving all rights to the work worldwide under copyright law. The license is especially used by database as it allows an effective removal of sui generis database right.

Three Levels of Accessibility

Creative Commons offer not only a choice of options but also of different levels of access. In contrast to the relatively heavy terms of other licenses, which require a full reinterpretation of the specialized legal text, CCs are divided into three levels of accessibility (Loren 2007):

- At the level of content itself (and its reworking and reuse). A small set of codified symbols indicates which option was chosen.
- These icons are linked to a simplified webpage that presents the terms of the license in a way that is easily understood by nonprofessionals (a kind of summary).
- Finally, based on this simplified webpage, users can access the detailed terms of authorization in the event of legal dispute.

International Goals Through Adaptation to National Legal Frameworks

While most licenses were exclusively designed for American legislation, internationalization was a priority in the creation of Creative Commons. Less than 4 months after the official announcement of its creation, the organization set up an office in Berlin to coordinate volunteer efforts to develop national versions. This was facilitated by preexisting transnational networks within the legal community, with 71% of its members from legal organizations (Dobusch and Quack 2008). In 2004, the license was transposed in 12 countries. The internationalization process continued at a steady pace until the end of the 2000s before slowing down. In 2016, the Creative Commons movement had representatives in 79 countries, with licenses adapted to 60 legislations. The complexity of the transposition process thus supplanted the original, nearly unattainable goal of a single, universal license.

Fifteen years after the first version (the project, launched in 2001, culminated in the publication of the first usable license and the organization's launch in 2002), CC licenses have crowded out most other free and acquired licenses in the cultural field, attaining a virtually monopolistic position.

A Large Offer Concentrated on Certain Platforms and for Certain Types of Content

In absence of an internal indexing system, statistical data on works, creators, and consumers of works is still quite patchy. Three types of sources can be used:

– Internal sources within the organization (such as State of the Commons or the Metrics Project). Since 2014, Creative Commons has

published a "State of the Commons" regarding uses on the main platforms (Flickr, YouTube, and Wikimedia), supplemented by a breakdown of off-platform web pages that use these licenses on Google (317 million pages in 2015). Previously, several statistical projects were done, including the article by Cheliotis et al. (2007) which marked the starting point.
- The data collected by certain Web actors, companies (e.g., Flickr and You Tube) or communities (e.g., Wikimedia Commons or Wikipedia).
- Occasional assessments from the scientific literature.

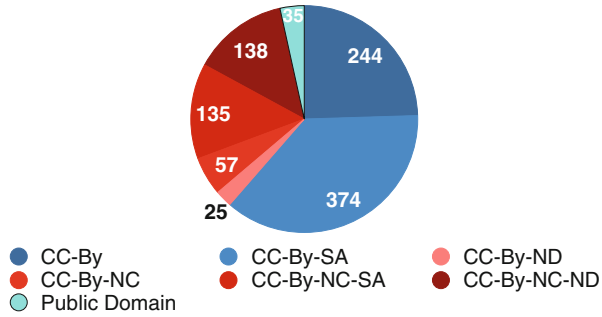
In the reviews conducted by the Creative Commons organization as well as researchers, the work is systematically equated to a "web page" with a valid link to a license. This semantic shift, though it considerably facilitates statistical collection, is nonetheless a simplification. Under this reserve, by 2015, there would have been 1.1 billion of what we will call "publications" under Creative Commons circulating on the web, a third on Flickr alone (350 million) and about 15% on various Wikimedia projects (140 million Wikipedia pages and 20 million images on Wikimedia Commons, the image library of Wikimedia projects).

Two periods stand out: a rapid take-off starting in 2004 (and particularly in 2005) and a period of stabile growth after 2010. In January 2007, there were 2.5 million publications (Kim 2007); 10 million by mid-February, 14 million by March, and 53 million by September. The end of the period covered by the Metrics Project on the contrary shows a relative stabilization, with 357 million publications in 2010 and 422 million in May 2011.

Although the movement is part of a widely internationalized approach, as Cheliotis et al. (2007) note, the use of CC tools reveals strong national disparities in the 33 countries where the licenses have been adapted. Economic development is the first factor that explains these differences; the magnitude of the digital divide has led to underrepresentation of China and South American countries compared to European countries and certain "highly-connected" Asian countries, such as Taiwan or South Korea. Social and cultural factors also play a role. For Cheliotis et al. (2007), the countries with the most widely distributed licenses and the most open options have a common profile: they are developed countries with strong digital infrastructures where the sharing of works, legal or illegal, is socially accepted (e.g., Sweden and Spain). In contrast, France, Belgium, and Italy form a kind of "Latin cluster," with a combination of strong license growth and more restrictive options. More recently, Latonero and Sinnreich (2014) found that growth was weaker in the United States than in Europe. The data collected in the 2015 edition of the State of the Commons on 317 million web pages also highlights the increasing internationalization of licenses: some countries in South America and Asia, such as Brazil and Thailand, have growth rates comparable to those in Europe. However, the distinctive characteristics of certain countries remain; Hungary, for example, has roughly one web publication per capita on Google (off-platform) under Creative Commons, versus only one in ten in France.

In the vast collection of publications under CC that exists today worldwide, certain publishing platforms clearly stand out: nine platforms publish 72% of the pages with link to a Creative Commons license, and only three of them, Flickr, Wikipedia, and Libre.fm, have 55% of all pages. The overall share of the eight platforms already enumerated in 2014 increased from 66% to 68% in 2015 (State of the Commons 2015).

The breakdown of licenses by type of activity also provides interesting information, with (in decreasing order) photographs (39 million), articles, stories and documents (47 million), videos (18 million), audio recordings (4 million), scientific articles (1.4 million), open educational resources (76,000), and an "other" heading (multimedia, 3D, etc.) of 23,000 (State of the Commons 2015). While the new growth drivers tend to be noncultural productions, scientific or educational, most notably, photographs and musical recordings dominate among cultural productions. The use of Creative Commons for the production of videos is only in its infancy (the CC-By license, for example, was only introduced



Creative Commons and Culture, Fig. 1 Breakdown of licenses in millions of publications. The most open licenses are indicated in shades of *blue* (From public domain to CC-By-SA). The most "closed" licenses are indicated in shades of *red* (From CC-By-NC-ND to CC-By-SA) (Source: Data form State of the commons 2015)

on YouTube in 2011) and has not yet been studied specifically.

In the field of photography, after a rapid take-off on Flickr between 2004 and 2006 (26 million photographs at the end of 2006 versus 300,000 in 2004), the number of new photographs under CC reached a plateau in late 2008 (39 million). Figure 1 remained steady at the end of 2009 and 2010 and then fluctuated sharply (24 million at the end of 2014 and 48 million at the end of 2015).

Wechsler (2010) observed the same pattern for music platforms. On Soundclick and Jamendo, initial growth was rapid until 2005–2006 and then stabilized. Certain music categories are also better represented: electronic music authors and classical music performers, respectively, represented 28% and 26% of audio publications under Creative Commons (Wechsler 2010). Concerning electronic music, a similar phenomenon was observed on Jamendo (Bazen et al. 2014). Two complementary explanations are put forth by the researchers. According to Wechsler, these two niche markets have fairly weak marketing prospects compared to rap or pop, for instance. Bazen, Bouvard, and Zimmermann insist, however, on the highly digitized production conditions of electronic music to explain this phenomenon.

Thus, the use of Creative Commons often reflects shared affinities or reluctance within a community of creators who, each in their own way, make strategic choices by opting for Creative Commons licenses and certain options more specifically.

A Continuum of Practices between Active and Passive Users

The very concept of users and free license users raises questions. A user is at once the "creator" who proposes original or derivative works under a CC license, the "consumer" who uses them or the platform that hosts them. However, few studies address these concepts in all their aspects.

Creators: To Each Their Choice

The studies available provide some information about creators who make their works available under Creative Commons licenses – a practice that is still marginal. For Latonero and Sinnreich (2014), only 5.3% of Internet users in the United States posted content under CC licenses. The overall figure for Internet users for the eleven other countries studied (the notion of contribution may, in this case, refer to creators of cultural works, wiki project contributors, etc.) was 17%. In a poll for Creative Commons (2009), only 6% of all American creators of original web publications had published content under CC in the last 12 months; among all American creators of web publications derived from original publications, only 3.3% of reworked publications had been done under CC in the last 12 months. The term user in this case does not refer to consumers of publications but to those who use them creatively (remixing, sampling, contribution. See the methodology presented in Creative Commons (2009, pp. 24–25)). Though being a minority in terms of the penetration rate among contributors, licenses

play an important role in some of them by meeting specific and changing needs.

Several studies emphasize the fact that creators of content under Creative Commons fall into a new intermediate category between amateurs and professionals. Wechsler (2010) notes that creators of music under CC are significantly less motivated by commercial gains than by social more social or political incentives (sharing of works, commitment to a change in intellectual property legislation, etc.).

Bazen et al. (2014) draw attention to close ties between amateurs and professionals on the Jamendo music platform. Here the majority of creators are amateurs, with professional musicians representing albeit a nonnegligible minority (22% among music groups and 18.5% among solo artists), contrary to the myth that Creative Commons only concern those who practice art as a hobby. The use of licenses reflects neither the dissolution of professional practices nor to the formalization of amateur practices, but rather an unprecedented equilibrium marked by the relaxing of the strict boundaries between "amateur" and "professional" (Bazen et al. 2014).

Beyond the hybridization of professional and amateur practices, the choices creators make from the different options available to them among CC licenses are not without consequence.

To clarify the more or less extensive nature of the various usage rights, a first indicator distinguishes the most open licenses, which allow for modification and commercial use, from the others. The ratio of "open" licenses to "closed" ones (noncommercial, no change possible) has increased considerably: according to internal Metrics data, between 2003 and 2010, after initial phase of fluctuation of 20% and 30%, the ratio jumped to 40% in 2010. The State of the Commons, which took over starting in 2014, shows that the most open licenses accounted for 56% of publications in 2014 and 65% in 2015. CC-By-SA and CC-By predominated here (with 374 and 244 million works in 2015, respectively). By also integrating works in the public domain, licenses allowing for commercial reuse and modifications thus represented almost two-thirds of the works identified (653 million out of 1 billion), a sharp contrast to a decade ago when the most open licenses were but a small minority. In the two cultural sectors where Creative Commons licenses are used most – photography and music – the preference today is for using the most open licenses.

In addition, Cheliotis (2009) notes recurring segmentation between two extremes: the first is characterized by the maintenance of purely artistic control over works through the use of noncommercial licenses, the second by a more altruistic relinquishing of rights (CC-By and CC-By-SA license), including for commercial uses. Bazen et al. (2014) also identified a divide within the population studied: a good quarter of creators intend to put up as few barriers as possible to the circulation of their work (BY), while half use more specific strategies. Cheliotis explains this divide as being driven by motives that are more pragmatic than ideological, given that creators with a low expectations with regard to the market are those most willing to opt for more open licenses. Wechsler (2010) reaches similar conclusions: among the creators interviewed, no self-proclaimed professional recommended the use of licenses that authorize commercial use. This population is therefore not representative of all Internet users who publish under Creative Commons licenses, nonprofessionals being more likely to use open licenses.

Beyond these divisions, a single creator often moves from one sphere to another depending on the situation: 33% of creators with more than one album to their credit on Jamendo used several CC licensing options (Wechsler 2010). Similarly, 50% of Soundclick's contributors publish their works based on classic copyright rules or using the various Creative Commons licenses depending on the case.

The Audience: An Unknown Variable

Most surveys on Creative Commons users focus on creators, which may seem paradoxical given that one of the objectives of these licenses is precisely to increase the diffusion of works. There is no evaluation/assessment of the overall audience of works under Creative Commons. In 2013, Wikimedia projects reached the

500 million mark for one-time visitors (per month), and Flickr the tens of millions (14 million for the United State alone). The number of one-time visitors is no longer calculated due to the audience's switchover to mobile devices (and the lack of identification measures to protect users' privacy).

These figures offer us a scale of magnitude but do not allow us to calculate a total audience, given the considerable overlap among one-time visitors between the sites and the fact that content under Creative Commons is designed to circulate freely. Google search results often include extracts from Wikipedia; photographs from Wikimedia Commons and Flickr are often used on press sites. A somewhat active user is or will be regularly exposed to content under Creative Commons without necessarily even knowing it.

Latonero and Sinnreich (2014) consider a specific audience segment, that of users who do targeted searches for content under Creative Commons or in the public domain, implying that they know of the existence of these licenses and their implications. The survey showed that 13% of Internet users in the United States belong to this "insider" audience segment, and that this figure jumps to 31% on average in the 11 other countries studied. Another study (Creative Commons 2009) focuses on an even more active and informed segment – that of users who not only consult works but use them to create new works (remixes, articles, etc.). A survey conducted by Greenfield for Creative Commons shows that 6% of respondents used them this way. In keeping with the hypothesis of cultural remix defended by Lawrence Lessig (2009), the population of active users overlaps with the American creator population also surveyed by Greenfield: only 28% had never published under Creative Commons licenses. The Creative Commons audience is thus a concentric circle with different levels of initiation in terms of licenses and different levels of use (from passive use to recreation).

The various licensing options are thus as many potential models of dissemination to suit the different individuals' pragmatic expectations.

Different Economic Models for Diverse Uses

The idea of a new form of collaboration between economic activities based on free culture has been widely discussed in the literature. Using software programs as support for his idea, Yochai Benkler (2002) theorized a cooperative alternative system called "commons-based peer production" specifically for the digital world. This model is a complement to those established since Ronald Coase's, the contract and the market. In this model, groups of individuals successfully collaborate on large-scale projects by following signals that are neither price- nor hierarchically-based. When it comes to producing information, this mode of production offers systematic advantages over the other two models; even though individuals do not directly reap the benefits of their participation in the collective project, their efforts have a greater impact than they would on the marketplace or in the firm, as commons-based peer production serves both to identify people best-suited to any given aspect of a project and to allocate resources to those able to put them to the best use according to optimal matching logic.

Beyond purely cooperative models based on volunteerism and free contributions, we find within the "free" world different business models of commercial service activity. Foray et al. (2006) underline that an understanding of the economics of open source should not be based on the marked dichotomy between analyses in terms of intrinsic and extrinsic motives, or in terms of community versus market-based economy. Rather, contemporary models of production and distribution of information products are hybrids models. "The hybrid economy "will dominate the architecture for commerce on the Web. It will also radically change the way sharing economies function. The hybrid is either a commercial entity that aims to leverage value from a sharing economy, or it is a sharing economy that builds a commercial entity to better support its sharing aims" (Lessig 2009, p. 177).

Market enterprises relying mainly on Creative Commons have had limited thus far. "The record labels supporting CC licenses are niche players. . .

Except for limited experiments, CC licenses have not yet been adopted by independent or major labels" (Wechsler (2010, p. 122). However, a hybrid economy has been created to support and/or complement business models in the digital economy. Digital technology has effectively led researchers to identify new business models that are better suited to the new conditions of production and distribution of information goods, particularly given that copyright laws are becoming almost impossible to enforce (Varian 2005). Two-sided market models based on advertising revenues occupy a key position in this economy. We find them in organizations that use CC licenses: 50% of Jamendo's advertising revenues (by number of pages viewed) were paid back to creators (Russi 2011). YouTube allows users to enhance content available under CC-By by associating it with advertisements. However, the perception of advertising revenue for covers is still complex: most intermediaries that register titles in the identification program of YouTube works (ContentID) do not accept works under free licenses: "Many of the online music distribution company. . . [do not accept them] into their distribution with YouTube's Content ID system." Although they sometimes exist, two-sided market models do not dominate in the CC license economy the way they do in the digital economy overall.

The literature on the specific subject of CC licenses reveals a multitude of economic models. To the extent that CC license users are both the creators and final consumers of works and platforms (see supra), whose aspirations are inherently heterogeneous, we propose three types of models that in each case highlight the choices of these three main categories of economic agents. However, these are only ideal types, whereas any given organization or economic agent typically combines the various models.

Fueling Celebrity Capital: The Choice of Still-Unknown Creators or Fan Communities

As mentioned above, CC licenses are widely used by a new intermediate category somewhere between amateurs and professionals. For many of them, greater publicity is more important than direct remuneration because the cultural economy is, in fact, an economy of brand recognition and reputation-building. A survey of artists who have opted for CC licenses (Wechsler 2010) shows that they often use them as a launch pad to create a buzz around their work: "Releasing music under a CC license is another way to get more publicity. By legalizing sharing, artists enable their fans to promote their music" (p. 129). "CC licenses may increase the diffusion of music and boost artists' reputations. These benefits may then be monetized in the form of more concerts and/or increased donations from fans" (p. 189).

CC are particularly suited to nonprofessionals who wish to make their work known, with little or no expectation of remuneration, and artists aspiring to become professionals. In this economy, symbolic remuneration – pride in collective participation within the community and recognition by peers – is strong. Becoming part a professional circuit can likewise be a powerful incentive for contributors, as drawing the attention of producers, employers, and public financial institutions in the hopes of funding future creations has become essential in an economy rich in cultural goods supply.

Communities of fans also form on the basis of sharing and celebrity. The Wikia website, created by Wikipedia cofounder Jimmy Wales, has become a major documentary reference in cultural industries. Collaborative encyclopedias fed by fans exhaustively cover certain fictional worlds, like that in Star Wars (160,000 articles put as on Wikipedia CC-By-SA license), and help build its reputation.

Harnessing End Users' Willingness to Pay

In so far as diffusion under CCs does not necessarily mean free access to works, a monetary contribution from end users may be requested. Certain forms of funding such as crowdfunding, which is based on prepurchase by consumers, are little used. Erickson et al. (2015) estimate the percentage of content under CC licenses on these platforms at 1%. End users of content under CC licenses are solicited mainly through voluntary donations or purchases of additional freedoms. On Magnatune and Jamendo, consumers can buy

music albums for the price they wish based on a "Pay what you want" formula (Russi 2011). For Wechsler (2010), consumers would be willing to pay an additional fee for works under Creative Commons licenses if they had the possibility of modifying them, for example. On the Beatpick platform, works are sold without DRM (Russi 2011). Others, like those devices used in the free software economy, for an additional fee offer the possibility of more open licenses by removing certain constraints, such as the obligation to use the same license for derivative works. In 2007, the Creative Commons organization designed a specific mechanism for managing the combining of several licenses: CCPlus, for example, is often used in the field of musical creation. A special symbol indicates that content users have the right to use content under a more open license (for example, allowing for commercial reuse): "The architecture of the CC-Plus scheme enables a commercial economy to co-exist with, or be grafted onto, the sharing economy created by the Creative Commons system" (Russi 2011, p. 106).

Adhering to the Values of a Community: The Choice of Institutions More So than that of Individuals

For Himanen (2001), contributors' motives for choosing free licenses hinge neither on a restrictive hierarchical structure nor on monetary incentives, but rather are based on a passionate relationship to works, flexibility of scheduling, appreciation for their "free" nature, and symbolic remuneration. Individual commitment is driven by an appreciation for the notion of give and take. In this model, there is no compulsory direct remuneration but rather remuneration through emotional dynamics and a relationship with the larger community.

Economic motives are sometimes secondary for creators. On Jamendo, although 22% chose CCs because the platform requires it, 20% did so because they saw them as a way to create a buzz, and more than 60% adhered to the idea of sharing that CC licenses promote (Bazen et al. 2014, p. 19). Furthermore, CC are particularly well adapted to scalable works, which are by nature perfectible, and can combine the works of different contributors. They were largely designed to facilitate the job for those who reuse and adapt the earlier versions of works of their colleagues more so than to market those of the original authors.

Beyond the individual choices of creators and/or end users, the use of CC licenses is by and large a choice made by organizations like Wikipedia and Flickr, which host, as we have seen, the vast majority of publications under CC license (see Supra) and impose more than propose the use of free licenses for works available on their sites. The hybrid models somewhere between market and cooperative coordination that are the strength of these organizations can also become Achilles's heel of this free culture economy by forcing the cohabitation of demands sometimes too contradictory and communities sometimes too different. As Lessig predicted: "No one builds hybrids on community sacrifice. Their value comes from giving members of the community what they want in a way that also gives the community something it needs" (Lessig 2009, pp. 177–178).

References

Bazen S, Bouvard L, Zimmermann J-B (2014) Jamendo et les Artistes: Un nouveau Modèle pour l'Industrie Musicale?

Benkler Y (2002) Coase's penguin or Linux and the nature of the firm. Yale Law J 112(Winter):369–446

Cheliotis G (2009) From open source to open content: organization, licensing and decision processes in open cultural production. Decis Support Syst 47(3):229–244

Cheliotis G, Chik W, Ankit G, Tayi KG (2007) Taking stock of the creative commons experiment monitoring the use of creative commons licenses and evaluating its implications for the future of creative commons and for copyright law. Singapore Management University

Creative Commons (2009) Defining "noncommercial" (Unsigned collective report)

Creative Commons (2015) State of the commons. https://stateof.creativecommons.org/2015/data.html

De Filippi P & Ramade I (2013) Les licences Creative Commons: Libre Choix ou Choix du Libre? In Christophe Masutti Camille Paloque-Berges Benjamin Jean (dir.), Histoire et cultures du Libre, Framabook, 2013, pp 341–378

Dobusch L, Quack S (2008) Epistemic communities and social movements: transnational dynamics in the case of creative commons. Social Science Research Network, Rochester

Erickson K, Heald PJ, Homberg F, Kretschmer M & Mendis D (2015) Copyright and the value of the public domain: an empirical assessment. Report commissioned by the Intellectual Property Office. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/561543/Copyright-and-the-public-domain.pdf
Foray D, Thoron S, Zimmermann J-B (2006) Open software: knowledge openness and cooperation in cyberspace. In: Brousseau E, Curien N (eds) Internet and digital economics. Cambridge University Press
Himanen P (2001) L'éthique hacker et l'esprit de l'ère de l'information. Exils, Paris
Kim M (2007) The creative commons and copyright protection in the digital era: uses of creative commons licenses. J Comput-Mediat Commun 13(1):187–209
Latonero M, Sinnreich A (2014) Tracking configurable culture from the margins to the mainstream. J Comput-Mediat Commun 19(4):798–823
Lessig L (2009) Remix: making art and commerce thrive in the hybrid economy. Penguin Books, New York
Lessig L (2001) The future of ideas: the fate of the commons in a connected world. Random House, New York
Loren LP (2007) Building a reliable semicommons of creative works: enforcement of creative commons licenses and limited abandonment of copyright. George Mason L Rev 14:271–328
Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge/New York
Russi G (2011) Creative commons, CC-plus, and hybrid intermediaries: a Stakeholder's perspective. Int Manag Rev 7:103
Turner F (2006) From counterculture to cyberculture – Stewart brand, the whole earth network, and the rise of digital utopianism. University of Chicago Press, Chicago
Varian HR (2005) Copying and copyright. J Econ Prospect 19(2):121–138
Wechsler J (2010) Openness in the music business – how record labels and artists may profit from reducing control. PhD dissertation, Technical University of Munich

Creativity

Christian Barrère
Faculty of Economics, Laboratoire R.E.G.A.R.D.S, University of Reims, Reims, France
ISMEA, Paris, France

Abstract

First of all creativity approach considers the creative behavior of judges and jurists in the legal processes. Afterwards it studies the specificities of legal tools, mainly IPR, regarding creative products.

Definition

From the 1960s, cognitive sciences and economics developed a new paradigm centered on creativity which no more dedicated to a marginal role, mainly in the artistic field, but appearing as a key component of human thinking and of social activity. For economists creativity is related to two conditions:

- It is a mental ability to create, i.e., to introduce a new thing (opus, idea, representation, etc.) where previously there was nothing similar (Sternberg 2006), as the image of God's creation.
- Goods based on creativity are mainly produced by the human brain, an idiosyncratic input, and do not result from the use of generic and standard economic resources as energy, equipment, and labor.

This new paradigm has two main consequences on the development of law and economics.

Firstly, it leads to change the basic model of the economic agent, from the standard homo oeconomicus, using his substantive rationality, to a creative person. That raises the issue of creativity in the making of law and in the evolution of legal systems. What are the ways according to which law is evolving and what is the role of creativity in this process? Is it playing differently in the common law and the civil law systems? Can the judicial decision be a creative one? Under what circumstances and within which limits?

Secondly, the creativity paradigm focuses on the evolution from an industrial society, based on the use of natural and human energy, to a creative economy based on creativity, on the use of the human brain and imagination (Potts 2011). The Lisbon Strategy, defined in 2007, decided to make European Union the most creative economy in the world. Thus, creativity has a strong value implying protection. Moreover, creativity appears as cumulative and noncumulative knowledge. Creativity close to science works as an input for the

production of new knowledge that is more developed, so that the old knowledge melts, over time, into the new one, which ceases to have a value as such. On the other hand, creativity close to arts gives noncumulative products that will not be replaced by better quality amenities in the future (Picasso is not a "better" painter than Poussin and Botticelli) and constitute creative heritages which are not fungible unlike the heritages of cumulative knowledge. So their value hugely matters. Thus, the protection of the value of creativity is a new key issue for economic development. Nevertheless, the specificities of creativity challenge the standard system of intellectual and industrial property rights:

- *Defining the entitlement*. The first problem in defining property rights is to identify all the resources (present and past including heritage), which currently have creative effects or can produce some effects in the future, to identify all the producers of resources, to separate their contributions, and to distribute rights among them so as to give each producer exclusive rights in their resources and control over the effect of their creative contribution. All that is often very difficult because creative production is frequently a team production and uses new and old creativity accumulated in heritages.
- *Organizing a market for IPRs*. In order to define property rights that can be transferred through a market process, resources have to be evaluated. But non-separability and nonadditivity, the dominance of creativity, and the difficulties of measurability disrupt the evaluation of the effects of the resources. Idiosyncrasy is a characteristic of creativity that makes it even more difficult to infer the value of used resources from the value of their effects.
- *Enforcing IPRs*. The third problem is to enforce the definition, entitlement, and transfer of property rights. Piracy and opportunistic behavior result from difficulties in identifying resources and in entitling them in order to define exclusive rights.
- *Justifying IPRs*. The difficulties to clearly define and valuate the productive resources and their holders imply difficulties to justify the present distribution of property rights. Normative problems arise. Is it fair to give the main part to the creator? Or to the owner of the firm? As usual in the field of intellectual property rights, the distribution of monetary earnings is far from contributing to human happiness or social development – is it fair that Einstein's income had been so small compared to that of Bill Gates?

Cross-References

► Innovation

References

Potts J (2011) Creative industries and economic evolution. Edward Elgar, Cheltenham

Sternberg RJ (2006) The nature of creativity. Creat Res J 18(1):87–98

Further Reading

Barrère C, Delabruyère S (2011) Intellectual property rights on creativity and heritage: the case of the fashion industry. Eur J Law Econ 32(3):305–339

Benghozi PJ, Santagata W (2001) Market piracy in the design-based industry: economics and policy regulation. Econ Appl LIV(3):121–148

Cohen JE (2007) Creativity and culture in copyright theory. UC Davis Law Rev 40:1151–1205

Fauchart E, von Hippel E (2008) Norms-based intellectual property systems: the case of French chefs. Organ Sci 19(2):187–201

Madison MJ, Frischmann BM, Strandburg KJ (2010) Constructing cultural commons in the cultural environment. Cornell Law Rev 95:657–710

Towse R (2001) Creativity, incentive, and reward: an economic analysis of copyright and culture in the information age. Edward Elgar, Cheltenham

Credibility

Jeong-Yoo Kim
Economics, Kyung Hee University, Seoul, Republic of Korea

Abstract

Two kinds of credibility will be distinguished, depending on the source of information. The

first credibility, which will be called type I credibility, is regarding information about the player's intention and the second credibility, which will be called type II credibility, is regarding information about the player's hidden characteristic. The former problem of credibility occurs when a player makes a promise or a threat (in a certain period of time) that he will do something in the future and then something happens between the two points of time. Then, the promise or the threat may not be credible, since the player will not have the incentive to keep it any more. The latter occurs when the informed player tries to pretend to be a better type by exploiting the uninformativeness of the other player. Then, the message conveying his information may not be credible as long as the incentive to pretend exists.

Definitions

Cheap talk	A costless, nonbinding, and nonverifiable message of a player.
Credible threat	A threat is credible when a player would find it in his interest to carry out the threat whenever he is called upon to do so.
Discovery	The legal process by which civil litigants are entitled to demand relevant evidence from each other.
Empty threat	A threat is empty if it is not credible.
Private information	Information possessed by fewer than all the players in a game.
Separating equilibrium	An equilibrium in which each possible type chooses a different action so that the type is revealed ex post.
Signaling	Choices by those who possess private information that convey information.
Subgame	Any part of a game that meets the following three conditions: (i) it starts from a single decision node, (ii) it includes all the nodes that follow the node, and (iii) if it includes a node in an information set, it also includes all other nodes in the information set.
Type	Hidden characteristic of a player.

Introduction

Information is widely dispersed throughout the economy, and its distribution is quite heterogeneous among economic agents. No one knows everything about the economy. Some agents may know better in a field, but others may know better in other fields. As such, there are informed players and uninformed players about almost anything in social interactions. Then, uninformed players may seek the information they need from the informed players, but the informed ones do not necessarily have an incentive to provide true information to the uninformed ones. For example, a mechanic may quote a fairly high price for a repair even though he found the problem only minor. A doctor may recommend his patients unnecessary medical tests and treatments. Presidential candidates often try to rope in votes by churning out meaningless pledges that are hardly feasible. Besides, it is not clear whether it is socially desirable always to be honest. As a typical example, it is controversial whether it is better to tell a patient of a cancer diagnosis. This draft addresses the issue of credibility, that is, when the messages of informed players are credible, when it is socially desirable to be credible, and what kind of legal and institutional devices are needed if they do not match, i.e., if informed players have an incentive not to be credible even if social efficiency requires credibility.

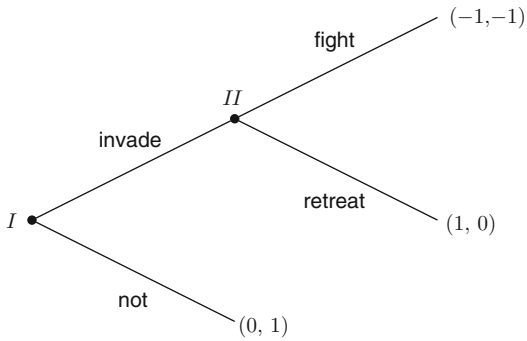
In this entry, I distinguish two kinds of credibility, depending on the source of information. The first credibility is regarding information about the player's intention (future choice) and the second credibility is regarding information about the player's hidden characteristic. The former is called the issue of dynamic inconsistency in literature. It occurs when a player makes a promise or a threat (in a certain period of time) that he will do something in the future and then something happens (new information is realized or the other player makes some decision) between the

two points of time. Then, the promise or the threat may not be credible, since the player will not have the incentive to keep it any more. The latter is called the issue of adverse selection. It occurs when the informed player tries to pretend to be a better type by exploiting the uninformativeness of the other player. Then, the message conveying his information may not be credible as long as the incentive to pretend exists. I will call the former type I credibility and the latter type II credibility.

This entry is organized as follows. In the next section, I explain the issue of credibility more formally by using the terminology of game theory and examples in general economics. In section "Applications to Law and Economics," I provide its applications to law and economics. Section "Conclusion" contains some future research directions and concluding remarks.

Where Credibility Comes From?

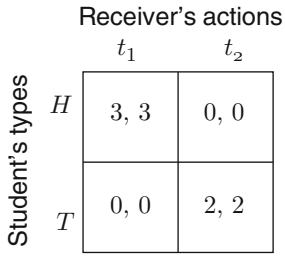
Consider the following simple game (Fig. 1). Two countries are in a war. Country 1 has two options: invade (I) or not (N). If country 1 invades, two options are available to country 2: either fight against it (F) or retreat (R). If country 1 does not invade, however, the game ends. As is well known, the game has two Nash equilibria (NE): (I, R) and (N, F). In particular, the second NE is noteworthy. "Not invade (N)" is optimal for country 1 given country 2 chooses to play F, since his invasion would entail "fighting (F)," whereas his opposite choice simply ends the game. In other words, country 1 finds invading against his interest because of country 2's threat of fighting. However, this strategy F involving the threat of fighting is never credible, although it constitutes NE. It could be credible if this game is played via a third-party referee who collects the strategies that each player submits to him and plays out the strategies by dictating them to play according to their submissions. In this scenario, once a player submits his strategy, say F, it is a commitment. He could not change it later, because the referee plays out the game. In this dynamic game, however, country 2 has a chance to change his strategy after the game begins. He can rethink at his decision node and change his mind after country 1 invades. Obviously, country 2 will prefer R to F, once country 1 invades. In other words, the strategy F which is a threat to fight if country 1 invades is not credible and thus called an "empty threat." The ex post optimal choice for country 2 when the other has already invaded is different from his ex ante optimal choice, because country 1's invasion has changed the situation. What country 2 should care about once invasion has occurred is a small subgame starting from his decision node, not the whole game. This dynamic inconsistency problem is quite universal. In dynamic games, a NE may involve a strategy which is not credible. Generally, a threat is credible if and only if it is a rational choice in every subgame, implying that it must survive backward induction. This argument is due to Selten (1965). Is there any way to make an empty threat credible? One could think of various commitment devices. One well-known example is to eliminate an option, for example, by burning the bridge (leading to the retreat).



Credibility, Fig. 1 Type I credibility

For the issue of the second kind of credibility, consider a game situation in which a player possesses private information. This is called a game of incomplete information. Suppose the informed player sends a message (regarding his private information) to the uninformed player and then the uninformed player responds to it. This special game of incomplete information is called a signaling game. In a signaling game, the informed player moves first, so his choice can be a signal of his private information, and the uninformed player may be able to infer the unobservable

information from observing the signal. Can the signal of the informed player be credible? Can it convey meaningful message regarding his private information? Of course, it can, if the signaling costs differ across types (i.e., values of private information). For instance, suppose that a firm wants to recruit workers of higher ability and more able workers can be educated at lower costs. Since a more able worker will try to signal his high ability by choosing to be more educated but a less able worker cannot do so, a higher education can be a credible signal for higher ability. This argument is due to Spence (1973). On the other hand, plain talks such as "I am a very able person" or "I am not able" involve no signaling cost so that the signaling costs of the two messages cannot differ for any type. So, the next question is whether costless signals such as talks can be credible at all. If the messages under consideration are "I am able" and "I am not able," the message of "I am able" can never be credible, because any worker would prefer this message to the other message, as long as both messages are equally cheap (costless) and both of them are taken seriously. However, what if the values of private information are not vertical but horizontal in the sense that the private information involves two kinds of ability that need nice matching between a firm and a worker, for instance, academic work and gardening? In this case, both a firm and a worker want good matching so that there are common interests. Then, a worker will not have an incentive to lie, even if academic work is more respected. For example, in a game illustrated in Fig. 2, the informed player who is good only at gardening will not say to the firm "I apply for academic work." Generally, if players have common interests, even cheap talk messages can be credible. This argument is due to Crawford and Sobel (1982).



Credibility, Fig. 2 Type II credibility

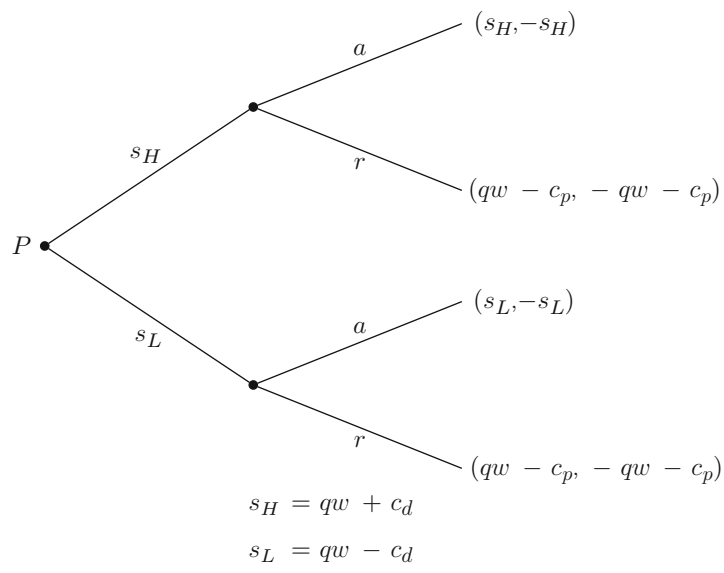
Applications to Law and Economics

The issue of credibility can occur in many legal situations in which a (potential) defendant (D) and a (potential) plaintiff (P) are involved with strategic interaction sequentially, but I focus only on civil/criminal procedures. I will use the following notation:

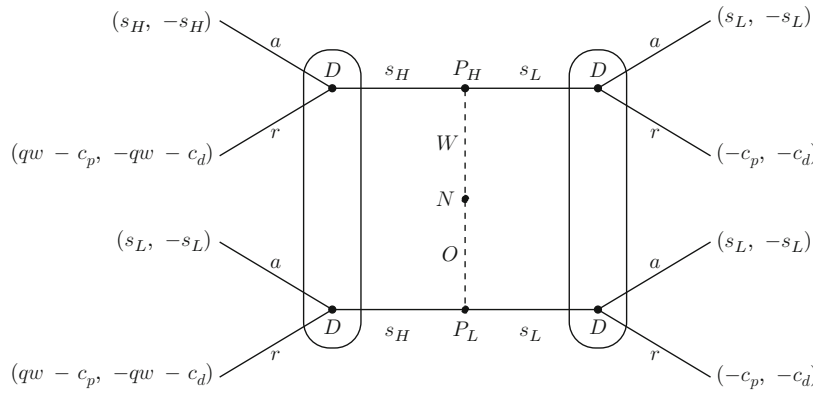
w = P's damage amount
q = D's liability (=P's winning probability at court)
c_p = P's litigation cost
c_d = D's litigation cost

Model 1 Consider a game illustrated in Fig. 3. This is a simple game of pretrial negotiation. P has only two options: either high settlement demand ($s_H = qw + c_d$) or low settlement demand ($s_L = qw - c_p$). If a settlement demand is rejected, they go to trial, causing each party to bear their respective litigation costs. This game has many NE. For instance, it is a NE that P makes a low demand s_L due to D's threat to reject any demand higher than s_L. In this NE, P makes a low demand s_L which is accepted by D for sure. However, D's threat is not credible, because D will accept a high demand any how if it is assumed that indifference in payoffs is resolved in favor of settlement. Thus, the only sensible outcome is that P makes a high demand which is accepted by D for sure. This is an example of type I credibility.

Model 2 The game provided in Fig. 4 is a modification of Fig. 3. Now, P's damage amount is his private information and is either w or 0. If the damage amount is 0 or more generally very low, the case is called frivolous suit. Since the informed party moves first in this game, it is a signaling game and the settlement demand made by P has a signaling effect. A high-type P - (of damage amount w) has no incentive to make a low demand, but a low-type P randomizes between a high demand and a low demand; in



Credibility, Fig. 3 Type I credibility in pretrial negotiation

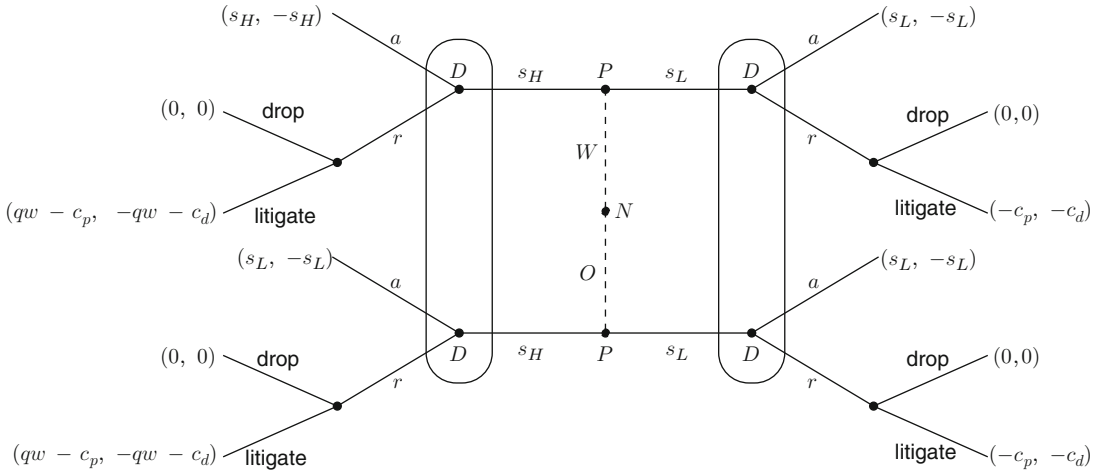


Credibility, Fig. 4 Type II credibility in pretrial negotiation

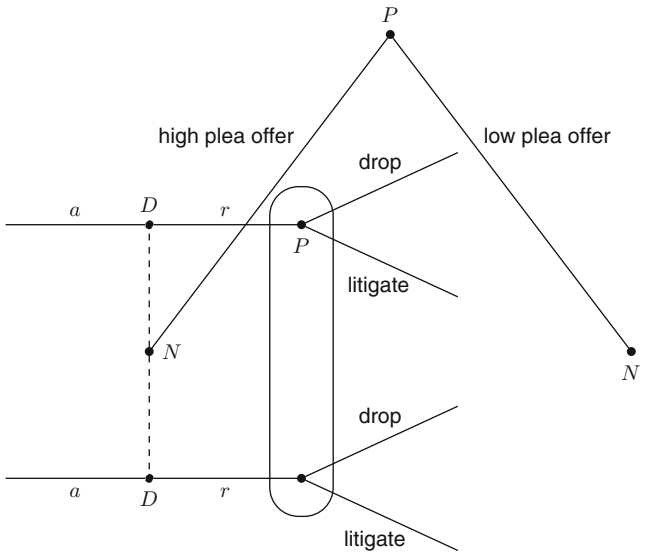
other words, he bluffs in order to extract a positive settlement amount at the risk of costly litigation. Thus, a high demand may not be credible in the sense that it could come from an undamaged P, while D can be sure that a low demand comes from a low-type P implying a frivolous suit. This is an example of type II credibility.

Model 3 The outcome of Fig. 4 relies on the implicit assumption that P commits to litigation when his demand is rejected. However, a low-type P has no reason to proceed to costly litigation, since his expected payoff at court is negative. He would rather drop the case once his demand is rejected. In other words, P's threat to go to trial if his demand is rejected is not credible. Taking this into consideration, I slightly modify Figs. 4 and 5 by incorporating P's option to withdraw the case. With the option, a low-type P always withdraws the case once his demand is rejected. This increases his expected payoff when he makes a high demand, which will make him choose a high demand with a higher probability. Both types of credibility are involved in this model.

Model 4 The issue of credibility that occurs when the option to drop the case is added becomes clearer in plea bargaining illustrated in Fig. 6. The defendant is either guilty (G) or innocent (I). The defendant knows his type G or I, but the



Credibility, Fig. 5 Both types of credibility in pretrial negotiation



Credibility, Fig. 6 Both types of credibility in plea bargaining

prosecutor does not. The prosecutor cares about both type I error and type II error. Grossman and Katz (1983) and Reinganum (1988) both identified a separating equilibrium in which a guilty defendant accepts the prosecutor's plea bargaining offer and an innocent one rejects. But both rely on a critical assumption that the prosecutor surely goes to trial once the defendant rejects his offer. In fact, it is not a credible threat to go to trial, because the prosecutor should know that the type who rejects his offer is innocent in the separating equilibrium. The prosecutor who cares about type I error as well as type II error should drop the case rather than proceed to trial. Grossman and Katz claim that a separating equilibrium cannot be obtained without the prosecutor's commitment to go to court. However, Kim (2010) shows that a (semi)separating equilibrium can be obtained without commitment power if mixed strategies are allowed. It can be a mixed strategy equilibrium that the defendant rejects the prosecutor's plea offer with some probability and the prosecutor chooses to litigate with some probability.

Model 5 Go back to Fig. 3 and consider a version of incomplete information game in which D has private information (about q) instead of P. In equilibrium, if P believes that D's liability is high, he will make a high demand which will be rejected with some positive probability. However, if D is actually a low type, he knows that they will end up with an inefficient outcome of costly trial after he rejects the high demand made by P. Is there any way for D to convince P that he is actually a low type? Suppose that D can use a cheap talk message before the game is played. Then, I can demonstrate that this pre-play communication can induce a more efficient outcome by avoiding some costly trial even though a plain talk is just a costless signal. Consider the following strategies. A high type of D announces "H" (i.e., "I am highly liable for your losses"), and a low type announces "L." P makes a high demand s_H if he observes the message "H" and makes a low demand s_L otherwise. The high demand is always accepted and the low demand is rejected with some probability α. This communicative outcome can be an equilibrium if $s_L = Lw + c_d$ and $s_L < s_H < Hw + c_d$. A high (bad) type of D cannot imitate a low (good) type by saying "L," since it might induce costly trial with some probability, so he would be better by losing s_H for sure rather than $Hw + c_d$ with probability $1-\alpha$ if α is low. In this cheap talk game, common interests between P and D (predisposition to avoid costly trial) enable costless communication messages to be credible. Credibility of cheap talk in pretrial negotiation is discussed in Kim (1992, 1996).

Disparity between the equilibrium outcome and the efficient outcome is obvious. As far as some information problem is present, the efficient outcome in which all legal disputes are resolved out of court could be hardly achieved even if various costly or costless signals are used. Is there any way to improve efficiency? If voluntary decentralized decisions cannot ensure efficiency, we could resort to institutional means such as binding contracts or law. For example, the government may adopt a rule imposing a penalty for misrepresenting information (e.g., mandatory discovery rules). Also, reputational concerns may alleviate the credibility problem.

Conclusion

The credibility issue is important in many legal situations beyond pretrial bargaining. For example, should we trust the announcement of the government to strictly regulate certain behavior and impose a severe penalty on it? Since regulation itself is costly to the government, it is optimal for the government that people believe the announcement of the government and discipline their behavior, but if the government knows this, it has no reason to enforce the costly regulation. It is dynamically inconsistent. As such, the government announcement need not to be credible unless the government cares about its long-term reputation or it is a legal commitment. I believe that readers can find more interesting examples in which a lack of credibility weakens the policy of the government.

References

Crawford V, Sobel J (1982) Strategic information transmission. Econometrica 50:1431–1451
Grossman G, Katz M (1983) Plea bargaining and social welfare. Am Econ Rev 73:749–757
Kim J-Y (1992) Does cheap talk matter in pre-trial negotiation? Seoul J Econ 5:301–315
Kim J-Y (1996) Cheap talk and reputation in repeated pre-trial negotiation. Rand J Econ 27:787–802
Kim J-Y (2010) Credible plea bargaining. Eur J Law Econ 29:279–293
Reinganum J (1988) Plea bargaining and prosecutorial discretion. Am Econ Rev 78:713–728
Selten R (1965) Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit. Zeitschrift fur die gesamte Staatswissenschaft 12:201–324
Spence M (1973) Job market signaling. Q J Econ 87:355–374

Credit Creation

▶ Credit Expansions

Credit Expansions

Juan Ramón Rallo
OMMA Center of Studies, Madrid, Spain

Abstract

Credit means deferred payment. Therefore, credit expansion means deferring more payments. The process of granting more credit inside the economic system can foster economic coordination, but it can also lead to intertemporal imbalances.

Keywords

Credit expansion; banks; interest rates; business cycle

Synonyms

Credit creation; Credit supply

Definition

Credit expansion is the process by which economic agents grant credit. Credit expansion can be either coordinated or uncoordinated.

Credit and Its Types

Credit (or debt) is the contractual right of an economic agent (the creditor) to receive a future economic good or service from another (the debtor). Though the nature of such good or service can vary, in monetary economies the repayment or valuation of credit in terms of money is increasingly common. In accounting terms, the monetary value of a debt is a *financial liability*.

Debts repayable in money arise when the debtor promises to deliver money to the creditor in the future. The source of that promise can be either:

1. A previous transfer of money from the creditor to the debtor, which gives rise to a loan, otherwise called *financial credit*. In accounting terms:

   or

2. A previous delivery of goods or services other than money from the creditor to the debtor, which gives rise to an outstanding account, otherwise called *trade credit*. In accounting terms:

Credit from a loan			
Creditor balance sheet at t = 1		**Creditor balance sheet at t = 2**	
Assets	**Liabilities + equity**	**Assets**	**Liabilities + equity**
Treasury: €10,000	Equity: €10,000	Credit: €10,000	Equity: €10.000
Debtor balance sheet at t = 1		**Debtor balance sheet at t = 2**	
Assets	**Liabilities + equity**	**Assets**	**Liabilities**
0	0	Treasury: €10,000	Debt: €10,000

Credit from an outstanding account			
Creditor balance sheet at t = 1		**Creditor balance sheet at t = 2**	
Assets	**Liabilities + equity**	**Assets**	**Liabilities + equity**
Commodities: €10,000	Equity: €10,000	Credit: €10,000	Equity: €10,000
Debtor balance sheet at t = 1		**Debtor balance sheet at t = 2**	
Assets	**Liabilities + equity**	**Assets**	**Liabilities + equity**
0	0	Commodities: €10,000	Debt: €10,000

The Rate of Interest

The common note of financial and trade credit is that both imply a deferred payment from the debtor and thus both are unsettled transactions: the debtor has a pending obligation to fulfill. As such, every credit involves two essential features: maturity and risk. All credit has a maturity because its repayment is due in the future; it also involves risk as future repayment depends on the ability and willingness of the debtor to fulfill its obligation.

Consequently, granting credit implies that the creditor accepts a deferred payment from the debtor, i.e., the creditor is willing to wait for and bear the risk inherent in future repayment. This is the reason why credit usually comes at a price. That price is called the rate of interest. The rate of interest can thus be understood as the difference in value between a spot, non-risky payment and a deferred, risky payment (Knight 1921; Fisher 1930).

Banks

As every credit transaction involves a particular interest rate, opportunities for arbitrage through intermediation will emerge. In every market where price differentials do appear, arbitrageurs can reduce them by taking a long position (buying side) in the relatively cheap good or asset and a simultaneous short position (selling side) in the relatively expensive good or asset (Fekete 1996).

Banks are the main arbitrageurs in the credit market. They go into debt at low interest rates in order to lend at higher interest rates. As intermediaries, banks specialize in granting credit by obtaining it, lending not their own funds but those of other economic agents. Therefore they do not grant credit out of nothing but out of their own liabilities and match every credit with a debit.

Bank balance sheet	
Assets	**Liabilities + equity**
Commercial credit: €100,000	Demand deposits: €100,000
Mortgage: €60,000	Covered bonds: €60,000
Buildings: €35,000	Equity: €40,000
Cash reserves: €5000	

Banks' intermediation serves a useful purpose in coordinating creditors and debtors. However, banks can also distort such coordination. Creditors and debtors have certain preferences about the maturity and risk they wish to assume. When they bargain directly with each other, they tend to reach an optimum agreement that matches both agents' needs. When, by contrast, they each bargain separately with a financial intermediary, the agreed sets of conditions need not be mutually compatible: banks can offer creditors maturity and risk conditions that are inconsistent with those offered to debtors. They might do so tempted by the profit opportunity implicit in the interest rate differential between usually low-interest, short-term, safer debts and usually high-interest, long-term, riskier debts.

Credit Expansion

The granting of new, previously inexistent credit is known as credit expansion (since credit is tantamount to debt, credit expansion could also be termed debt expansion). A *coordinated* credit expansion occurs when the sets of conditions offered to creditors and debtors are fully compatible. An *uncoordinated* credit expansion, by contrast, occurs when the sets of conditions offered to both agents are *not* fully compatible.

Coordinated Credit Expansion

Coordinated credit expansions occur when risk and maturity conditions offered by financial intermediaries to lenders and borrowers coincide. This implies that financial intermediaries aim to match the maturity and risk of their assets and liabilities, bringing into mutual consistency their cash outflows and inflows. In other words, banks only grant long-term loans when they hold enough long-term financing sources and only grant loans to risky borrowers when their creditors have willingly acquired risky bank liabilities (such as subordinated bonds).

Since credit was defined as the contractual right of a creditor to receive an economic good or service from a debtor, coordinated credit expansions imply that financial intermediaries can modify in the aggregate neither the future preferences for goods and services among lenders nor the availability of those future goods and services among borrowers. Therefore, in the aggregate, preference for resources among lenders and availability of resources among borrowers should be matched in their maturity and risk profiles. Historically, the prescription for matching banks' assets and liabilities has been known as the *golden rule of banking* (Hübner 1854).

The most common bank liability is the demand deposit (or, alternatively, the banknote). A demand deposit is a debt callable at sight by the creditor. Due to its immediate and safe convertibility into money, demand deposits are generally used as money substitutes: they can be endorsed to third parties as payment in a given transaction. Economic agents' demand for cash balances is partly satisfied by the supply of banks' demand deposits, thereby transforming hoarding into a means of funding new credits through banking (Selgin 1988). However, if banks wish to expand credit by adequately matching the preferences of both lenders and borrowers, the assets backing demand deposits should be highly liquid, i.e., easily convertible into cash.

Some authors have gone as far as to defend *full-reserve banking* as the only banking practice that can adequately coordinate the preferences of creditors and debtors. Under full-reserve banking, demand deposits can only be issued against an equal sum of cash reserves (Fisher 1935; Friedman 1960). In other words, the reserve ratio (cash reserve divided by demand deposits) must be 100 %. In fact, those authors usually refer to the *money multiplier* as those demand deposits issued in excess of cash reserves. For instance, in the following example, the reserve ratio is 20 % (cash reserves amount to 20 % of all demand deposits) and, therefore, the money multiplier is 5 (the money multiplier is inverse to the reserve ratio).

Bank balance sheet	
Assets	**Liabilities + equity**
Commercial credit: €90,000	Demand deposits: €100,000
Cash reserves: €20,000	Equity: €10,000

Nonetheless, other economists recognize the possibility of backing demand deposits with other assets of practically equal liquidity such as short-term trade credit (Adam Smith 1776; Melchior Palyi 1936). If financial intermediaries could not issue demand deposits against assets other than cash, then demand deposits would no longer be proper instruments for financial intermediation, i.e., for coordinating creditors and borrowers: they would just become custody deposits (Huerta de Soto 1998).

There certainly seems to be some scope for banks to fund their expansions of short-term credit by increasing their demand deposits without undermining the coordination among lenders and borrowers. However, funding long-term and risky credits with demand deposits (or other kinds of current liabilities) would lead banks to an uncoordinated credit expansion.

Uncoordinated Credit Expansion

Uncoordinated credit expansions occur when risk and maturity conditions offered by financial intermediaries to lenders and borrowers do not coincide. This implies that financial intermediaries mismatch the maturity and risk of their assets and liabilities, throwing their cash outflows and inflows into mutual inconsistency. In other words, banks engage in long-term lending by issuing short-term debt or invest in high-risk assets by issuing debt to risk-averse savers.

Bank balance sheet	
Assets	**Liabilities + equity**
Mortgages: €100,000	Demand deposits: €100,000
Junk bonds: €50,000	Senior debt: €40,000
	Equity: €10,000

The problems of uncoordinated credit expansion have been generally recognized: (1) banks become exposed to bank runs unless their liabilities are protected by deposit insurance schemes and by institutional lenders of last resort (Diamond and Dybvig 1983); (2) the financial system as a whole becomes fragile and unstable despite the existence of the previously mentioned fire walls (Fekete 1983; Minsky 1986); (3) the real economy runs the risk of entering a business cycle due to a provision of credit for investment that is far in excess of real savings (Mises 1912; Hayek 1931).

The law of large numbers might make it easy to think that an individual bank can avoid the adverse consequences of its own asset and liability mismatching. However, this is certainly not possible for the whole banking system. A widespread uncoordinated credit expansion endogenously gives rise to the financial instability and business cycle that ultimately undermine the very operation of the law of large numbers (Huerta de Soto 1998).

Financial regulation is, in fact, increasingly acknowledging the dangers involved in maturity and risk mismatching. Basel III, for instance, requires banks to keep a minimum capital buffer in order to absorb potential losses (their common equity and Tier I capital should at least be equal to 4.5 % and 6 %, respectively, of their risk-weighted assets). Moreover, Basel III imposes liquidity constraints: the Liquidity Coverage Ratio (banks must have enough liquid assets to cope with 1 month of net cash outflows) and Net Stable Funding Ratio (aimed at ensuring that banks hold a minimum amount of stable funding based on the liquidity characteristics of their assets and activities over a one-year horizon). Other authors, however, suggest that precisely the absence of government intervention would result in a competitive environment conducive to a natural and much more effective self-regulation (Selgin and White 1994).

Conclusion

Whichever the proposed solutions to avoid them, it is clear that uncoordinated credit expansions tend to distort a financial system. Therefore, they should not be confused with *coordinated* credit expansions, which, by matching creditors and debtors' preferences, increase the number of economic exchanges in a sustainable manner.

Cross-References

- ▶ Banks
- ▶ Hayek, Friedrich August von

References

Diamond D, Dybvig P (1983) Bank runs, deposit insurance, and liquidity. J Polit Econom 91(3):401–419
Fekete A (1983) Borrowing short and lending long. Committee for Monetary Research and Education, Charlotte
Fekete A (1996) Towards a dynamic microeconomics. Laissez-Faire n°5. Universidad Francisco Marroquín, Guatemala, pp 1–14
Fisher I (1935) 100% money: designed to keep checking banks 100% liquid; to prevent inflation and deflation; largely to cure or prevent depressions; and to wipe out much of the national debt. The Adelphi Company, New York
Fisher I (1930) The theory of interest. The Macmillan Company, New York
Friedman M (1960) A program for monetary stability. Fordham University Press, New York
Hayek F (1931) Prices and production. Routledge and Sons, London
Hübner O (1854) Die Banken. Hübner, Leipzig
Huerta de Soto J (1998) Dinero, crédito bancario y ciclos económicos. Unión Editorial, Madrid
Knight F (1921) Risk, uncertainty, and profit. Houghton Mifflin, Boston
Minsky H (1986) Stabilizing an unstable economy. Yale University Press, New Haven
Mises L (1912) Theorie des Geldes un der Umlausfsmittel. Duncker and Humblot, Munich
Palyi M (1936) Liquidity minnesota bankers association. Minneapolis
Selgin G (1988) The theory of free banking. Rowman& Littlefield, Totowa
Selgin G, White L (1994) How would the invisible hand handle with money. J Econ Lit 32(4):1718–1749
Smith A (1776) The wealth of nations. W. Strahan and T Cadell, Scotland

Credit Supply

▶ Credit Expansions

Credit Trading

▶ Emissions Trading

Credit: Rating Agencies

Alessandro Romano
China-EU School of Law, China University of Political Science and Law, Beijing, China

Abstract

Credit rating agencies (CRAs) are pivotal players in financial markets, and in fact their conduct has attracted the attention of scholars, media, and policy analysts. A very common claim is that CRA behavior contributed to the

explosion and the propagation of the recent financial crisis. This entry sketches the functioning of the market for ratings and explores the market failures by which it is characterized. Moreover, this entry briefly presents some of the proposals advanced by the law and economics literature to induce CRAs to issue accurate ratings.

Introduction

The main activity of credit rating agencies (CRAs) is providing investors and regulators with certifications of the quality of financial assets. Any lender is interested in knowing what is the likelihood that the borrower will honor his debt, and ratings serve exactly this need. In a nutshell, ratings are informed opinions on the probability that the lender (issuer of the financial asset) will not repay the borrower (investor). In other words, ratings are an estimate of the probability of default (PD) of a given bond. In some cases, ratings also account for other factors, like the expected magnitude of the losses associated with a possible default of the issuer (loss given default (LGD)). When ratings are not accurate, they can be either inflated or deflated. A rating is inflated when the CRA overestimates the creditworthiness of the rated bond. Instead, a rating is deflated when the creditworthiness of the issuer is underestimated. CRAs can potentially be useful actors on financial markets because they help reducing information asymmetries between the issuers of the rated assets and regulators and investors. However, over the last decade CRAs have been at the center of a very heated debate both at the policy level and on the scientific literature as, allegedly, they have played a crucial role in the recent financial crisis. In particular, many scholars and policy makers have argued that CRAs issued inflated ratings thus contributing to the explosion and the propagation of the financial crisis (White 2010). This entry investigates how much truth there is behind these accusations, what are the reasons that might have lead credit rating agencies to inflate their ratings, and what are the proposals advanced by the law and economics literature to induce CRAs to issue accurate ratings. Two preliminary caveats are required. First, ratings can be divided in two broad categories: solicited ratings and unsolicited ratings. Solicited ratings are requested by the issuer that pays a fee to be rated by the CRA and provides the CRA with relevant information. Instead, unsolicited ratings are spontaneously issued by the CRA that does not receive any fee and are generally based on information available to the public. Because solicited and unsolicited ratings are associated with drastically different incentives for the parties involved, these two kinds of ratings cannot be analyzed together. This entry focuses only on solicited ratings as the fees collected issuing this kind of ratings generate the vast majority of CRAs revenues. Second, not all solicited ratings are equal. Ratings of structured finance products are markedly different – and way more problematic – than ratings of corporate bonds.

History of the Market for Ratings

In 1909 John Moody published the first publicly available bond ratings. Other firms soon engaged in this practice creating the market for ratings. At the time, these ratings were sold to investors who paid to have an overview of the creditworthiness of a number of issuers. This is a business model currently referred to as investor-pays model. However, a number of reforms and technological innovations completely transformed the landscape of the market for ratings. A detailed analysis lies beyond the scope of this entry, but a quick overview of the most relevant changes is useful to understand what the problems in the market for ratings are.

Starting from the 1930s, regulators began to grant credit rating agencies an increasingly fundamental role in the functioning of financial markets. For example, in 1936 bank regulators issued a decree that was aimed at preventing banks from investing in "junk bonds," as defined by "recognized ratings manuals." As the only recognized ratings manuals were Moody's, Poor's, Standard, and Fitch, regulators de facto gave to the judgments of these four rating agencies (later to

become three when Standard and Poor's merged into Standard & Poor's) the status of law (White 2010). Insurance and pension regulators adopted very similar rules (White 2010). In the 1970s, Moody's, Standard & Poor's, and Fitch relevance on financial markets was further increased by the combined effect of two decisions of the Security and Exchange Commission (SEC). On the one hand, the SEC imposed that the minimum capital requirements of brokers-dealers had to be tied to the riskiness of their asset portfolio, and ratings were to be used to determine the level of risk. On the other hand, afraid that smaller CRAs not constrained by reputational concerns could issue inflated ratings to attract customers, the SEC dictated that only ratings issued by "nationally recognized statistical rating organization" (NRSRO) could influence the minimum capital requirements. Moody's, Standard & Poor's, and Fitch were the only CRAs that obtained the status of NRSRO. Combined with a number of other regulations, the effect of these reforms was to grant Moody's, Standard & Poor's, and Fitch a quasi-regulatory power.

Another piece of the puzzle is the change in the business model that CRAs undertook in the 1970s. Potential free riding problems associated with the diffusion of fast photocopy machines undermined the investor-pays model. In particular, because it was becoming easy, quick, and cheap to disseminate information, rating agencies feared that the content of their ratings could have circulated also among investors that did not pay the relative fee. Therefore, instead of exacting a payment from investors who wanted to see the ratings of a given issuer, CRAs started to require a fee from the issuer that they had to rate. Nowadays, 95% of CRAs revenues derive from the fees collected from issuers (Partnoy 1999). Last, over the last decades rating agencies started to rate structured finance products that were becoming more and more complex.

Summarizing, three main factors characterized the evolution of the market for ratings:

- An increased regulatory relevance of ratings that granted CRAs a quasi-regulatory power.
- The adoption of an issuer-pays model.
- CRAs started rating complex structured finance products.

Failures in the Market for Ratings

In the wake of the crisis, referring to CRAs the Nobel Prize winner Paul Krugman wrote that:

It was a system that looked dignified and respectable on the surface. Yet it produced huge conflicts of interest. Issuers of debt could choose among several rating agencies. So they could direct their business to whichever agency was most likely to give a favorable verdict, and threaten to pull business from an agency that tried too hard to do its job. (Krugman 2010)

To put it differently, CRAs rate their clients, and hence they could be inclined to cater to the needs of the latter to attract more business (Darcy 2009). This view has been extremely influential in the literature, in the policy debate, and in the media, but it overlooks the insights of reputational capital theory (e.g., Choi 1998). According to this theory, "the only reason that rating agencies are able to charge fees at all is because the public has enough confidence in the integrity of these ratings to find them of value in evaluating the riskiness of investments" (Macey 1998). Therefore, absent other market failures, reputational sanctions would discipline CRAs' behavior preventing them from inflating ratings. In this vein, the literature has looked beyond the conflict of interest and identified three other market failures that altered the functioning of the market for ratings.

First, reputational capital theory is grounded on the idea that investors are sophisticated enough to determine when ratings are inflated. However, if a large enough fraction of investors is Naive and cannot identify inaccurate ratings, reputational sanctions become largely ineffective (Bolton et al. 2012). Second, CRAs collect their fee only when they publish the ratings, and hence issuers could contact multiple rating agencies and request publication only for the most favorable rating received (Dennis 2009). This practice of shopping for the most favorable rating can result in rating inflation, especially for complex assets (Skreta and Veldkamp 2009). Third, as high ratings are

associated with regulatory benefits, issuers might be interested in purchasing good ratings, regardless of whether investors trust the ratings. Thus, when the regulatory benefits attached to high ratings are sufficiently relevant, a rating agency "finds it profitable to stop acquiring any information and merely facilitates regulatory arbitrage through rating inflation" (Opp et al. 2013).

Therefore, there is a combination of market failures that, associated with the issuer-pays model, induces CRAs to inflate their ratings. And indeed, while the effective contribution of rating inflation to the financial crisis is still disputed (Gorton and Ordoñez (2014) argue that the contribution was likely to be limited), a large part of the literature finds that ratings, especially of structured finance products, were inflated (e.g., Calomiris 2009).

Fixing the Market for Rating

The issuer-pays model, combined with the possibility of shopping for the most favorable rating, the regulatory benefits attached to high ratings, and the naivety of some investors, creates incentives for the credit rating agencies to inflate their ratings. As there is such a complex web of market failures, inducing CRAs to issue accurate ratings is no easy task. Moreover, not all ratings are equal, and ratings of complex structured finance products create more concerns than the traditional ratings of corporate bonds. The main reasons are that (i) structured finance products are more complex and rating inflation can be more severe for complex bonds (Skreta and Veldkamp 2009) and (ii) many structured finance products behave as economic catastrophe bonds, that is, these financial assets are less resistant to economic downturns and their defaults are highly correlated (Coval et al. 2009). Any proposal that aims at improving the functioning of the market for ratings should account for these differences.

The market for ratings is an oligopoly with high barriers to entry dominated by Moody's, Standard & Poor's, and Fitch, and hence an obvious solution to ameliorate CRAs incentives could be increasing the competition in the market (Hill 2003). Nevertheless, empirical research shows that under the *status quo*, competition *worsens* the quality of ratings (Becker and Milbourn 2011). In presence of rating shopping more competition negatively affects the quality of ratings because the issuer has more choices when searching for the most favorable rating (Becker and Milbourn 2011). Alternatively, as regulatory benefits attached to high ratings neutralize – or at least reduce the impact of – reputational sanctions, one obvious solution to improve the quality of the ratings is reducing their regulatory relevance (Flannery et al. 2010). The Dodd-Frank Act takes exactly this path, but it seems unlikely that the implemented reforms will suffice to eliminate both direct and indirect regulatory benefits derived from relying on ratings (Hill 2010). And indeed, the European regulator explicitly remarked that there are no perfect substitutes for ratings, and hence ratings are bound to have some regulatory value (Pacces and Romano 2015). As noted by Coffee (2011), regulators decided to assign regulatory value to ratings because they have limited information and cannot develop reliable measures of risk. In other words, ratings are a precious component of regulation, provided that they are accurate, because there is no guarantee that alternative solutions to identify excessive risk will not prove to be even more problematic. Therefore, reforms should attempt to improve CRAs' incentives while preserving – at least partially – the role played by ratings in financial regulation. Another possible reform is forcing CRAs to abandon the current issuer-pays model in favor of different business models (Mathis et al. 2009) or even introducing some sort of public funding for CRAs (Listokin and Taibleson 2010). However, the information contained in ratings has the nature of a public good because it can easily be disseminated among investors, and hence alternatives to the issuer-pays model are generally considered unworkable (Partnoy 1999; Coffee 2011). Another proposal is to pay rating agencies with the debt that they rate (Listokin and Taibleson 2010). In this vein, CRAs would be punished when issuing overoptimistic ratings because they receive debt that is worth less than they claim. Last, a path that has been widely

explored by the law and economics literature is to make the liability threat faced by CRAs issuing inaccurate ratings more credible. In fact, for many years CRAs have been de facto immune to liability claims (Coffee 2006; Partnoy 2006), and in the United States they were even put under the umbrella of the First Amendment on the freedom of speech (Deats 2010). Generally, the literature considers a negligence rule as the most appropriate to induce credit rating agencies to issue accurate ratings. Under this rule, rating agencies are asked to compensate the investor only when they have been negligent in formulating their rating. In Europe, the United States, and Australia the liability of CRAs is largely based on this logic. There are, however, a number of problems with this approach (Pacces and Romano 2015). First, it is extremely hard for courts to identify the optimal level of care (Coffee 2004), because ratings are complex and prospective judgments that necessarily involve at least some subjectivity on the part of the raters. Determining when there was negligence in the formulation of this prospective judgment has been defined a Serbonian Bog by the literature (Coffee 2004). Imprecise and uncertain standards of care are notoriously associated with an increase in transaction costs and more unpredictability of courts' behavior. Second, CRAs are not responsible for all the losses associated to a default. For example, CRAs did not lower Enron rating below investment grade until only a few days before the bankruptcy, thus fueling accusations of negligent behavior on their part (Frost 2007). However, while CRAs can be held liable for not detecting Enron's problems, they are certainly not liable for the fact that Enron went bankrupt. Therefore, they should be asked to compensate only a fraction of the harm associated with Enron's bankruptcy. Identifying which fraction of the harm is attributable to CRAs conduct is extremely challenging, if not impossible (Pacces and Romano 2015). If the liability threat is reinforced via a negligence rule, the combined effect of these two problems might be that rating agencies become exceedingly conservative in their judgments or even refuse to rate risky securities. These problems would be especially severe because, on the one hand, risky assets are exactly those for which ratings are more needed. On the other hand, CRAs can play a beneficial role only if they issue accurate ratings, not if they issue deflated ratings. Empirical evidence and theoretical studies suggest that this risk is concrete. The Dodd-Frank Act significantly increased CRAs' liability exposure, and Dimitrov et al. (2015) found that as a consequence the informative content of corporate ratings further worsened. At a theoretical level, Goel and Thakor (2011) show that when the liability threat is severe, credit rating agencies have an incentive to issue deflated ratings, because it is unlikely that courts will hold them liable for conservative ratings. To put it differently, the expected liability faced by CRAs issuing inflated ratings is larger than that faced by CRAs issuing deflated ratings.

The problems of a strict liability rule are as severe. On the one hand, if CRAs are asked to cover all the losses whenever an issuer they rated defaults, they would bankrupt almost immediately as the default of a single large issuer might cause losses that exceed the assets of a CRA. Moreover, under a strict liability rule, the injurer (here the CRA) acts as a de facto insurer (Priest 1987). It is common wisdom that only uncorrelated risks can be insured (Priest 1987), whereas defaults – especially of structured finance products – are highly correlated and concentrate during economic crises (Coval et al. 2009). Pacces and Romano (2015) attempt to cope with this problem proposing a less intrusive form of strict liability that relies mainly on market forces. Introducing a damage cap based on objective factors and corrections to shield CRAs from the risk of correlated defaults, Pacces and Romano (2015) argue that a modified regime of strict liability might induce CRAs to issue ratings that are as accurate as the available forecasting techniques allow.

In conclusion, the market for ratings is characterized by multiple market failures, and hence the literature is still struggling to find effective solutions.

Future Research

How to prevent future malfunctioning in the market for ratings is still an open issue. In the coming

years quantitative studies might attempt to disentangle the size of the impact of each market failure and to understand how these market failures interact with each other. For example, regulatory benefits and investors' naivety reduce the effect of reputational sanctions, but how their effects are related is more obscure. The relationship between the effects of these two failures might be additive (e.g., if each reduces the magnitude of reputational sanctions by 10%, then the joint effect is 20%), or the effects of the two market failures could stand in more complex relationships (e.g., the presence of naïve investors magnifies the effect on reputational sanctions of the regulatory benefits, thus producing a joint effect larger than 20%). And indeed, we have seen that the issuer-pays model in itself is not necessarily problematic, but it creates perverse incentives when combined with other market failures. A clearer and more quantitative understating of the interactions among market failures in the market for ratings might help improving the regulatory regime of rating agencies.

Another important question that awaits an answer is to which extent regulatory reliance on rating agencies is motivated. In turn, answering this question implies that alternative solutions and their possible limitations are explored. Flannery et al. (2010) attempt exactly this task, but more studies are warranted.

Cross-References

- ▶ Conflict of Interest
- ▶ Market Failure: Analysis
- ▶ Strict Liability Versus Negligence

References

Becker B, Milbourn T (2011) How did increased competition affect credit ratings? J Financ Econ 101:493

Bolton P, Freixas X, Shapiro J (2012) The credit ratings game. J Finance 67:85

Calomiris CW (2009) A recipe for ratings reform. Econ Voice 6:1

Choi S (1998) Market lessons for gatekeepers. Northwest Univ Law Rev 92:916–919

Coffee JC Jr (2004) Gatekeeper failure and reform: the challenge of fashioning relevant reforms. Boston Univ Law Rev 84:301

Coffee JC Jr (2006) Gatekeepers: the professions and corporate governance. Oxford University Press, Oxford

Coffee JC Jr (2011) Ratings reforms: the good, the bad and the ugly. Harv Bus Law Rev 1:231

Coval J et al (2009) The economics of structured finance. J Econ Perspect 23:3

Darcy D (2009) Credit rating agencies and the credit crisis: how the issuer pays conflict contributed and what regulators might do about it. C Bus Law Rev 2:605

Deats C (2010) Note, talk that isn't cheap: does the first amendment protect credit rating agencies' faulty methodologies from regulation? Columbia Law Rev 110:1818

Dennis K (2009) The rating game: explaining rating agency failures in the buildup to the financial crisis. Univ Miami Law Rev 63:1111

Dimitrov V et al (2015) Impact of the Dodd-Frank act on credit ratings. J Fin Econ 115:505

Flannery MJ et al (2010) Credit default swap spreads as viable substitutes for credit ratings. Univ Pennsylvania Law Rev 158:2085

Frost CA (2007) Credit rating agencies in capital markets: a review of research evidence on selected criticisms of the agencies. J Account Audit Finance 22:469

Goel AM, Thakor AV (2011) Credit ratings and litigation risk. Available at http://ssrn.com/abstract51787206

Gorton G, Ordoñez G (2014) Collateral crises. Am Econ Rev 104:343

Hill CA (2003) Rating agencies behaving badly: the case of Enron. Conn Law Rev 35:1145

Hill CA (2010) Justification norms under uncertainty: a preliminary inquiry. Conn Insur Law J 17:27

Krugman P (2010) Berating the raters, N.Y Times. http://www.nytimes.com/2010/04/26/opinion/26krugman.html?_r=0

Listokin Y, Taibleson B (2010) If you Misrate, then you lose: improving credit rating accuracy through incentive compensation. Yale J Regul 27:91

Macey (1998) Wall street versus main street: how ignorance, hyperbole, and fear lead to regulation. Univ Chicago Law Rev 65:1487

Mathis J, McAndrews J, Rochet J-C (2009) Rating the raters: are reputation concerns powerful enough to discipline rating agencies? J Monet Econ 56:657

Opp C et al (2013) Rating agencies in the face of regulation. J Financ Econ 108:46

Pacces A, Romano A (2015) A strict liability regime for rating agencies. Am Bus Law J 52:673

Partnoy F (1999) The Siskel and Ebert of financial markets?: two thumbs down for the credit rating agencies. Wash Univ Law Q 77:619

Partnoy F (2006) How and why credit rating agencies are not like other gatekeepers. In: Fuchita Y, Litan RE (eds) Financial gatekeepers: can they protect investors? Nomura Institute of Capital Markets Research/ Brooking Institution Press, Tokyo/Washington, DC, p 61

Priest GL (1987) The current insurance crisis and modern tort law. Yale Law J 96:1521
Skreta V, Veldkamp L (2009) Ratings shopping and asset complexity: a theory of ratings inflation. J Monet Econ 56:678
White LJ (2010) Markets: the credit rating agencies. J Econ Perspect 24:211

Crime and Punishment (Becker 1968)

Jean-Baptiste Fleury
THEMA, University of Cergy-Pontoise, Cergy-Pontoise, France

Definition

Gary Becker's 1968 "Crime and Punishment: An Economic Approach" is one of the first papers using economics to address the questions of crime and law enforcement. To Becker, crime generates costs to society, but fighting crime is also costly. There is, therefore, an optimal amount of crime which minimizes society's total loss and which can be attained by setting the optimal levels of punishment and probability of apprehension and conviction. From that analysis, Becker further claims that the role of criminal law and law enforcement policies should be limited to the minimization of society's loss. Crime is, therefore, framed as an external effect, and criminal law's purpose is redefined as the activity of assessing the harm incurred by crime in order to enforce optimal compensation.

Becker's 1968 "Crime and Punishment: An Economic Approach"

Gary Becker's 1968 "Crime and Punishment: an Economic Approach" is one of the first articles by a modern economist (post-World War II) to address crime (see also Eide et al. 2006). Due to its huge influence not only in creating the whole subfield of economic analyses of criminal behavior and public law enforcement (see the extensive survey of Polinsky and Shavell 2000) but also in the development of Richard Posner's economic analysis of law (see Posner 1993), the paper is worth studying. The paper is quite typical of Gary Becker's approach to economics. First, it applies standard microeconomic tools to the problem of public law enforcement, which was, up to 1968, a topic traditionally considered to be outside of the domain of economics. Although Cesare Beccaria and Jeremy Bentham interpreted criminal law in a utilitarian and economic framework during the eighteenth and nineteenth centuries, their works were progressively relegated to the fringes of both economics and criminology during the twentieth century and are generally mentioned only for historical references. Second, the scope of the paper is much broader than crime and law enforcement: it is a "generalization of the economist's analysis of external harm or diseconomies" (Becker 1968, p. 201), which offers to redefine both crime and criminal law along the lines of economic efficiency.

The Basic Model

Let's consider the first aspect of Becker's analysis: the application of standard microeconomic analysis to the question of law enforcement. To address the problem, Becker adopts an approach reminiscent of welfare economics and resorts to a social welfare function which computes the total social net loss generated by criminal activities, with such activities being defined – so far – exogenously by the law. To simplify the analysis, Becker reduces the scope of the social loss function to losses in real income only. The function can be decomposed into three different sections: first, the total net social damages due to criminal activity (D). Second, the costs of apprehension and conviction of offenders (C). Finally, the social cost of punishment (S). All these subpart depend on the number of offenses that are perpetrated during a given period of time.

Thus $L = D + C + S$.

Becker takes a utilitarian point of view: a given offense generates damages to the victim and society but generates gains to offenders that have to be

taken into account. Following the general economic assumptions, the additional offense yields positive and increasing harm to society and positive but decreasing gains to offenders. The production of law enforcement is an activity which is costly. Thus, if more resources are invested in increasing the probability of conviction *p*, the cost increases, and the marginal costs also increase. If the number of offenses increases, the cost of law enforcement increases, as well as its marginal costs. Finally, most forms of punishment (with the notable exception of fines) generate costs to society: prisons need personnel and facilities, while imprisonment also incurs costs to the offender depending on his opportunity cost of time.

Since all the subparts of the equation depend on the number of offenses, Becker's central point is to assess the optimal number of offenses in a given society. Clearly, crime not only generates costs to society: apprehending and punishing offenders also generate costs, so zero crime appears socially inefficient. There must be a total number of offenses that yields a socially desirable outcome. As Becker (1968, p. 170) famously wrote, the main question of the analysis is "how many offenses *should* be permitted, and how many offenders *should* go unpunished?" Public policy exerts an influence on the number of offenses, through two main variables. One is the probability of apprehension and conviction, *p*. The second is the level of punishment, *f*. To complete the model, thus, Becker analyzes the supply of offenses by criminals, which mostly depends on those two variables.

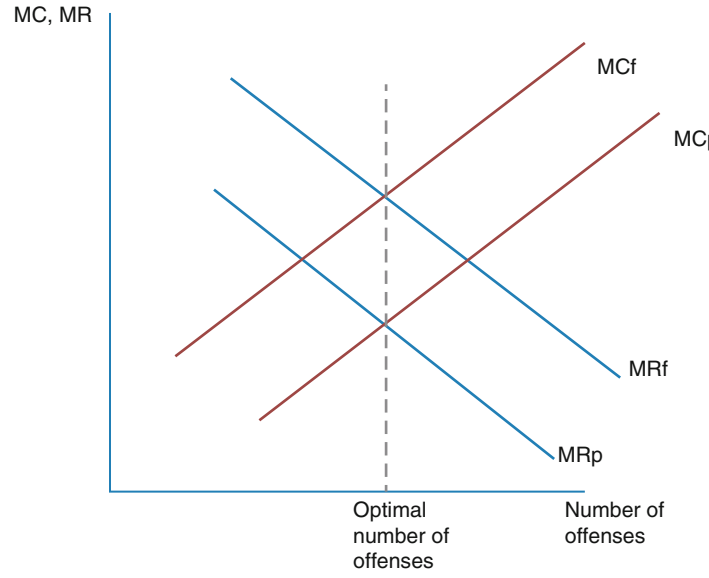
Although Becker's analysis of criminal behavior is perhaps the most remembered contribution of his 1968 paper, the actual model of individual behavior is relegated in a footnote, as Becker's central focus is on the market supply of offenses. Becker assumes that individuals are perfectly rational: they maximize the expected utility they derive from the net gains acquired from a criminal activity. It depends, therefore, on *p* and *f*. An increase in the probability of conviction *p* would reduce the expected utility from criminal activity, as would an increase in the level of punishment *f*. Becker's analysis shows that criminals tend to be risk lovers, which means that a given increase in *f* perfectly compensated by a decrease in *p* would leave the expected gains the same but would make criminals better off. Therefore, when criminals are risk lovers, the elasticity of supply of offenses with respect to *p* is greater than the elasticity of supply of offenses with respect to *f*.

C

Optimality Conditions

With the model now complete, Becker's formulation of the optimality conditions is organized so as to compare marginal costs and marginal benefits from a given increase in the quantity of offenses, generated by a decrease either in *p* or *f* or both. Regarding marginal costs, a small decrease in *f* would yield additional offenses, incurring therefore additional costs from damages and from conviction and apprehension (curve MCf in Fig. 1). The same could be said for a small decrease in *p,* but the marginal cost to society would be smaller than in the previous case because, this time, society saves some costs related to law enforcement: reducing the probability of cleared cases means roughly less resources spent on law enforcement (curve MCp). Both curves are upward sloping due to the increasing marginal damages and marginal costs functions associated to an increase in the number of offenses.

Marginal benefits are related to the social costs of punishment: more offenses means less punishment, and, therefore, less social costs due to these punishments. The curve is downward sloping. Therefore, marginal benefits are related essentially to two elements. The first is the coefficient transforming a given punishment into the social costs of such a punishment. The second is the elasticity of supply of offenses with respect to *p* and *f*. Indeed, the social "benefits" coming from an increase in offenses depend partly on the sensitivity of offenders to decreases in *p* or *f*. Because marginal cost must equal marginal revenue and because the marginal cost of a reduction in *p* is lower than the one due to a reduction in *f*, the marginal revenue related to a reduction in *p* has to be lower than the marginal revenue related to a reduction in *f* (the MRf curve is higher than the MRp curve). Becker shows that this is



Crime and Punishment (Becker 1968), Fig. 1 Optimality conditions

only possible if the elasticity of supply of offenses with respect to p is greater than the one with respect to f, in other words, if criminals react more strongly to an increase in p rather than f, that is, if they are risk lovers.

When marginal costs equal marginal revenue, one finds the optimal number of offenses, and the optimal values of p and f leading to such offenses.

From this analysis of optimality conditions, we can firstly conclude that Becker's paper provides a normative analysis of how to allocate resources in order to minimize society's loss, that is, to reach the optimal amount of crime by playing on the values of p and f. But, in the meantime, it also provides an evaluation of current and past policies against crime. Becker generally concludes that what can be observed in terms of actual probability of conviction as well as severity of punishment (and their evolution in time) is overall compatible with his normative statements. Therefore, in a sense, the theory provides also a positive analysis of how society considers crime and has responded to crime over time. Becker's comparative statics are in this respect enlightening, and two examples will be explored.

First, suppose that a given increase in crime yields higher marginal damage to society. The marginal cost curve would shift upward, and the optimal amount of crime would go down. This would be achieved by an increase in both p and f. Becker's crude empirical investigation showed, indeed, that the more serious the felony is, the more likely the criminal is going to be convicted and the harsher the punishment he will face. Second, suppose that a reduction in the marginal revenue came from an increase in the elasticity of supply (Ef) with regards to f – the social cost of punishment is therefore reduced, which means that the marginal revenue that society gains through additional offenses diminishes – then the optimal level of crime is reduced and is achieved through an increase in f. Such a result leads Becker to show that society would minimize its loss by engaging in price discrimination: different groups of offenders with different elasticities should be charged different levels of punishment. Becker shows that this is consistent with actual practice, where groups being insensitive to punishments, such as impulsive murderers or juveniles, generally face lower punishments and more therapy for similar crimes.

Criminal law and the General Theory of External Effects

Becker's economic analysis supports a view of criminal law and law enforcement practices

grounded on the maximization of social welfare. From that perspective, the main role of punishment is to compensate for the marginal harm done to victims and society. Becker considers other motives such as revenge and deterrence, as secondary. Thus, the amount of punishment does not depend on the specific conditions of the criminal, but on the marginal damage suffered by society, which needs to be compensated. Fines are the punishment that yields almost no social cost: contrary to imprisonment, fines are simply performing wealth transfers. They stand as the perfect compensation mechanism and should be, therefore, used whenever necessary.

From this normative analysis, Becker broadens the scope of his paper. First, he redefines completely the notion of crime, not as an exogenous activity deemed illegal by law, but as any activity which generates harm that was not compensated. Eventually, there are no differences between a person buying a car and a thief stealing one and compensating society afterward through a well-calculated fine. Only when the damage cannot be compensated by fines should the "debtor" repay the involuntary "creditor" as well a society through other forms of punishments, such as prisons. Second, he redefines criminal law: "the primary aim of all legal proceedings would become the same: not punishment or deterrence, but simply the assessment of 'harm' done by defendants," in order to calculate the levels of compensation (Becker 1968, p. 198). Thus, "much of traditional criminal law would become a branch of the law of torts, say, 'social torts', in which the public would collectively sue for 'public harm'" (Becker 1968, p. 198). Note that this aspect of the paper had a tremendous influence on Posner's subsequent economic analysis of law and his 1985 definition of crime as "market bypassing" (see Posner 1985, 1993). Consequently, Becker concludes that his analysis of crime is a generalization of the theory of external effects, placing criminal law and law enforcement as the central institutional setting to articulate a sort of Pigovian taxation. To Becker, crime becomes nothing more than an external effect which, if negative, has to be taxed and reduced to an optimal level. But Becker even considers the case of positive external effects, which should be regulated with subsidies, rewards, and other forms of cash prizes, which magnitude and probability to award would be the social variables to be controlled. Law enforcement, in this case, would have to spend resources to find and award inventors and other producers of external benefits.

Cross-References

- ▶ Becker, Gary S.
- ▶ Cost of Crime
- ▶ Criminal Sanctions and Deterrence
- ▶ Economic Analysis of Law
- ▶ Law and Economics
- ▶ Law and Economics, History of
- ▶ Posner, Richard

References

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217
Eide E, Rubin PH, Shepherd JM (2006) Economics of crime. Now Publishers Inc., Delft
Polinsky M, Shavell S (2000) The economic theory of public enforcement of law. J Econ Lit 38(1):45–76
Posner RA (1985) An economic theory of the criminal law. Columbia Law Rev 85(6):1193–1231
Posner RA (1993) Gary Becker's contributions to law and economics. J Leg Stud 22(2):211–215

Crime, Attitude Towards

Elena D'Agostino[1], Emiliano Sironi[2] and Giuseppe Sobbrio[1]
[1]Department of Economics, University of Messina, Messina, Italy
[2]Department of Statistical Sciences, Catholic University of Milan, Milan, Italy

Abstract

Crime negatively affects social welfare and reduces citizens' trust in public institutions and society. Perhaps the best starting point is trying to look at criminal behavior as human

and social phenomenon in order to understand what pushes people to illegal (and certainly risky) activities. The psychological literature has emphasized the role of attitudes as one of the main determinants of human behavior. We reviewed this literature and its applications to crime.

The Theory of Planned Behavior

Crime negatively affects social welfare and reduces citizens' trust in public institutions and society. Reducing crime is therefore a priority of every public agenda across the world, although what is the most effective program remains an open question. Perhaps, the best starting point is trying to look at criminal behavior as human and social phenomenon in order to understand what pushes people to illegal (and certainly risky) activities.

One approach to crime consists of viewing it simply as a human behavior. As such, the decision to commit a crime is the (more or less) logical consequence of a mental process involving morality, civic awareness, and personal character. According to Ajzen (1991) and his theory of planned behavior, an individual's decision to engage or not engage in a given behavior is anticipated by the formation of positive intentions toward that behavior. Intentions depend primarily on three variables: (1) personal *attitudes* toward a given behavior, (2) people's belief concerning whether he or she is expected by other individuals to perform that behavior (*subjective norms*), and (3) the individual's perceived ease or difficulty of performing the behavior (*perceived behavioral control*).

Applications

The theory of planned behavior has been tested in many fields through cross analysis of attitudes and behaviors. For example, environmental economics suggests that sustainable behaviors are widely promoted by positive attitudes, but the lack of accessible recycling infrastructures and other constraints work against the original intention. Health economics has used TPB to determine the obesity factors of overweight among Chinese Americans (Liou and Bauer 2007).

Testing the TPB model in a similar way in relation to crime is difficult because people are clearly not keen to admit they have committed crimes, so there is a lack of information concerning the last (and most important) part of the story, that is, intentions or behaviors. Nevertheless, if it is accepted that criminal behavior is a decision-making process, it is possible to understand the main factors that influence people's future illegal actions by focusing on attitudes, used to denote their internal evaluations of the extent to which respecting laws will have positive or negative consequences for their lives and happiness.

There are many attempts in the literature to analyze and understand people's attitudes toward legality and therefore toward crime. Using data from the world value survey, Torgler and Schneider (2007) investigated the determinants of attitudes toward paying more or less in taxes from a cross-country perspective, considering the impact of both sociodemographic and cultural background. Further studies have focused on the relationship between education and crime, commonly arguing that education and the associated higher earnings negatively affect both crime (among others, Buonanno and Leonida 2006) and psychological attitudes toward crime (Arrow 1997). However, these findings are far from conclusive. Groot and Van Den Brink (2010) found evidence of a relationship between high education levels and attitude toward serious crimes in the Netherlands. D'Agostino et al. (2013) confirmed these findings in a cross analysis involving most European countries.

Empirical studies show a strong positive relationship between fear of crime and media consumption (among others, see Barille 1984).

Concluding Remarks

What is common to all these papers is that each of them focuses on specific drivers for determining

attitudes toward crime and punishment (media, economic recession, education). Some of these contributions examines whether attitudes toward crime are influenced more by individual variables, involving personal features (such as gender, race, education, family), or by contextual variables, referring to institutional and economic aspects (such as corruption, GDP, growth, interpersonal safety). The results, although in line with the hypothesis of a combined effect of the context and of individual features, are someway counterintuitive: often, more educated individuals seem also to be more tolerant with criminals.

This result is probably due to the high heterogeneity of the class of criminal acts and to the different opinion with respect to the role of punishment and detention. More educated people are also those that more frequently deal with white-collar crimes, a sort of crime that is more widespread and tolerated in the upper class than assaults and physical violence.

This raises important questions about the social dimension of crime as a complex phenomenon involving the individual's life within society. On the other hand, more educated people tend to give the punishment a role that is more rehabilitative than punitive; all these aspects may explain why education moderates attitudes toward crime. However, with respect to crime, the link between attitudes and behavior remains under-investigated in comparison with other social objects. Some questions are still open. In more detail, it would be questioned whether education and wealth work as a consciousness that is an alternative to law to establish what is right and what is not right. On the other hand, does direct experience of corruption and poverty make people desperate to recognize law and legality as the only hope for a better life? And, more importantly, will people behave according or against to their attitudes? What seems to emerge from the literature is that attitudes are probably not enough to understand and to predict future behaviors but may work as an important starting point for any serious analysis of this phenomenon.

References

Ajzen I (1991) The theory of planned behaviour. Organ Behav Hum Decis Process 50:179–211

Arrow K (1997) The benefit of education and the formation of preferences. In: Behrman J, Stacey N (eds) The social benefits of education. University of Michigan Press, Ann Arbor, pp 11–15

Barille L (1984) Television and attitudes about crime: do heavy views distort criminality and support retributive justice? In: Surette R (ed) Justice and the media: issues and research. Charles C. Thomas, Springfield

Buonanno P, Leonida L (2006) Education and crime: evidence from Italian regions. Appl Econ Lett 13:709–713

D'Agostino E, Sironi E, Sobbrio G (2013) The role of education in determining the attitudes towards crime in Europe. Appl Econ Lett 20:724–727

Groot W, Van Den Brink HM (2010) The effects of education on crime. Appl Econ 42:279–89

Liou D, Bauer KD (2007) Exploratory investigation of obesity risk and prevention in Chinese Americans. J Nutr Educ Behav 39(3):134–141

Torgler B, Schneider F (2007) What shapes attitudes toward paying taxes? Evidence from multicultural European countries. Soc Sci Q 88(2):443–470

Crime, Incentive to

Derek Pyne
Thompson Rivers University, Kamloops, BC, Canada

Abstract

The standard economic model of crime assumes criminal incentives depend on (1) the expected payoff from crime, (2) the penalty if apprehended and convicted, (3) the probability of apprehension and conviction, and (4) attitudes toward risk. At times, penalties for failed criminal attempts are also taken into account.

Definition

Criminal incentives either motivate individuals to commit crimes or motivate them to abstain from crime. The key positive incentive to commit crime is the expected benefit a criminal receives from the crime. Key negative incentives involve deterrence

from the probability of detection and the resulting penalty. Increases in the opportunity cost of committing crime rather than engaging in regular employment also act as a negative incentive.

Criminal Incentives

The most common behavioral assumption in economics is that economic agents respond to incentives. Without this assumption, the economics of crime would not exist as an area of study.

The simplest model of a risk-neutral criminal's objective function has the following form:

$$U = B - qD \tag{1}$$

where B represents the benefits of crime, q represents the probability of apprehension and conviction, and D represents the disutility of the penalty if convicted.

B may represent either psychological or financial benefits. The entry by Leroch (2014) deals with a similar distinction between expressive crime and instrumental crime. The entry by Schneider and Meierrieks (2014) discusses the nonfinancial benefits of Terrorism. As those entries focus on expressive crime, this entry will concentrate on crimes with more tangible benefits.

If a criminal has the choice between working and committing crimes, B may be measured relative to his income from employment. If so, B may be a function of employment prospects, the minimum wage, and the criminal's education. Both a theoretical and an empirical finding is that unskilled crimes (e.g., violent and property crime) are negatively related to education and white-collar crimes are positively related to education (Lochner 2004). This is consistent with more educated individuals having greater opportunities (and hence incentives) for white-collar crimes. However, their higher market incomes result in lower benefits, relative to opportunity costs, for other crimes.

B also depends on the opportunities available. These opportunities decrease as electronic transactions replace cash (Wright et al. 2014). This may be a reason for recent falling crime rates.

The disincentive of punishment (D) can include imprisonment, fines, forfeiture, stigma, poor human capital accumulation, and other penalties. The entry by Prescott (2015) surveys a range of criminal sanctions and their implications for deterrence. The entry by Di Vita (2015) contains a discussion of sanctions and deterrence implications for environmental crime. The entry by Vannini et al. (2015) does the same for kidnapping. The entry by Donohue (2014) focuses on the specific case of capital punishment.

Poor criminals may not have the resources to pay fines. If so, fines are less of a deterrent than imprisonment. The opportunity cost of imprisonment will partly depend on forgone wages. Thus, the disincentive effects of both imprisonment and fines may be positively related to income.

With incarceration, poor prison conditions increase D and discourage crime (Katz et al. 2003; Pyne 2010). However, there is evidence that incarcerated criminals are more likely to reoffend when prison conditions are poor (Drago et al. 2011). This may be because harsher prison conditions lead to a deterioration of human capital and poorer future employment prospects. In other words, a reduction in prison conditions may lead to a contemporaneous increase in D while increasing future values of B (relative to the opportunity cost of working instead) for those incarcerated.

With incarceration, there is also evidence of an intertemporal relationship between B and D through prison increasing criminal capital (Bayer et al. 2009).

The stigma of conviction increases D. Stigma effects may work through social ostracism or poorer future employment prospects. Typically, it is assumed that stigma effects are greatest for the first conviction and then decline. If so, ceteris paribus, a criminal's incentives to commit crime increase after a first conviction. This is one explanation for the observation that most formal penalty schedules involve harsher penalties for repeat offenders. The increase may be necessary to counter the decreasing deterrent effects of stigma (Miceli and Bucci 2005).

Somewhat similar to stigma effects are effects on human capital. There is evidence that incarceration decreases the ability of juveniles to acquire human capital. If so, their wages as adults are lower, which lowers their future opportunity cost of incarceration. Pyne (2010) offers this as an explanation for the more lenient treatment of juveniles in most justice systems.

A wide range of other penalties can result in disutility. Examples include impaired drivers being barred from driving, licensed professionals losing their license, problems crossing borders due to convictions, and having to compensate victims.

The probability (q) of being caught and convicted for a criminal act provides a negative incentive to commit crime. This may be a function of a criminal's ability (Miles and Pyne 2015), reporting by victims and witnesses (Allen 2011), law enforcement expenditures (Vollaard and Hamed 2012), prosecutor behavior (Entorf and Spengler 2015), certainty of conviction (Entorf and Spengler 2015), and the standard of proof required for convictions (Pyne 2004).

Empirical evidence suggests that for many crimes, q is small. For example, based on data from a British longitudinal survey, Farrington et al. (2006) find that when assaults are excluded, on average there were 14 self-reported crimes for every conviction. When assaults are included, the ratio of self-reported crimes to convictions is 22.

Higher-ability criminals face a lower probability of being convicted for their crimes. This gives them a greater incentive to commit crimes (This assumes they are not also proportionally better at earning income from noncriminal activities.). Thus, it is possible that those with more convictions may have committed proportionately more crimes than those with fewer convictions (Miles and Pyne 2015).

In some cases, there may be free riding by enforcers which results in a less than optimal q. This, and related enforcement problems, are discussed in the entry by Hallwood and Miceli (2014) on modern piracy.

Whether the probability of being wrongly convicted of a crime changes the disincentive effects of q is an unsettled question. Many argue that the likelihood of being wrongly convicted of a crime reduces the net increase in the probability of being convicted of a crime when one actually commits the crime (Pyne 2004; Polinsky and Shavell 2007; Rizzolli and Stanca 2012). However, some have argued to the contrary. For example, Lando (2006) argues that one may be wrongly convicted of a crime committed by someone else, independently of whether one has committed a crime of their own.

Attitudes towards risk affect the relative deterrent effects of penalties and conviction probabilities. Equation (*1*) assumed individuals were risk-neutral. Thus, changes in penalties should have the same deterrent effect as changes in conviction probabilities that have the same effect on qD. If criminals are risk-averse, changes in penalties should have greater deterrent effects than changes in conviction probabilities. In either case, if enforcement and punishment is costly, it would be best to have punishments that do not have to be enforced (Becker 1968). The basic economic model of crime implies that if it were feasible to set penalties high enough, all crime could be deterred.

Not only are extreme punishments not used in practice but empirical evidence finds that changes in the probability of conviction have a greater deterrent effect than changes in penalties (Eide 2000). Several approaches have been offered to reconcile the basic model with the empirical evidence.

One approach involves explanations consistent with standard expected utility theory. For example, criminals may be risk lovers (Becker 1968). However, this is contrary to standard economic assumptions used in other contexts. Alternatively, Pyne (2012) shows that low-ability criminals may learn more about their innate ability from increased enforcement than from increased penalties, which reveal no information about ability. Another approach is to relax the standard assumptions of the expected utility model (Neilson and Winter 1997).

A different approach is to entirely reject expected utility theory. An example is

Al-Nowaihi and Dhami's (2010) use of composite cumulative prospect theory.

Another complication of the basic model involves the possibility that a criminal's attempt at a crime is unsuccessful. This will, in part, depend on resistance and precautions taken by the victim. Allen (2011) surveys literature related to victim behavior and offers an analysis.

The lower the punishment for unsuccessful attempts, the greater incentive criminals have to target victims who take fewer precautions. Ben-Shahar and Harel (1996) argue that this may be efficient if victim precautions are too high. Those who would otherwise exercise a high level of precaution have less incentive to do so as criminals are targeting those taking a lower level of precautions instead. However, if high-income potential victims can afford greater precautions, distributional considerations may also be involved.

The probability of success may depend on competition from other criminals. In some instances, imperfect enforcement may actually benefit some criminals by reducing competition from competitors (Miles and Pyne 2014).

Expected punishment also affects the criminal's choice between crimes. If the punishment for less socially harmful crimes increases, criminals have an incentive to substitute towards more socially harmful crimes (Stigler 1970).

There has been an increasing focus on behavioral economic considerations of criminal incentives. Typically, behavioral considerations have more of a quantitative effect on incentives than a qualitative effect. For example, hyperbolic discounting changes the value a prospective criminal places on a marginal increase in prison sentences, but the increase is still an incentive to not commit crime. Other behavioral considerations include loss aversion, issues involving probabilities, bounded rationality, and emotional considerations. For surveys of the behavioral economics literature on crime, see Garoupa (2003) and van Winden and Ash (2012). For a general introduction to its use in law and economics, see the entry by Ko (2014).

Cross-References

- ▶ Becker, Gary S.
- ▶ Behavioral Law and Economics
- ▶ Cost of Crime
- ▶ Crime: Expressive Crime and the Law
- ▶ Crime: Organized Crime and the Law
- ▶ Crime and Punishment (Becker 1968)
- ▶ Criminal Sanctions and Deterrence
- ▶ Death Penalty
- ▶ Environmental Crime
- ▶ Piracy, Modern Maritime
- ▶ Ransom Kidnapping
- ▶ Terrorism

References

Al-Nowaihi A, Dhami S (2010) The behavioral economics of crime and punishment. Discussion papers in economics, University of Leicester

Allen DW (2011) Criminals and victims. Stanford University Press, Stanford

Bayer P, Hjalmarsson R, Pozen D (2009) Building criminal capital behind bars: peer effects in juvenile corrections. Q J Econ 124(1):105–147

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217

Ben-Shahar O, Harel A (1996) The economics of the law of criminal attempts: a victim-centered perspective. Univ Pa Law Rev 145(2):299–351

Di Vita G (2015) Environmental crime. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York

Donohue JJ (2014) Death penalty. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York

Drago F, Galbiati R, Vertova P (2011) Prison conditions and recidivism. Am Law Econ Rev 13(1):103–130

Eide E (2000) Economics of criminal behavior. In: Bouckaert B, De Geest G (eds) Encyclopedia of law and economics. Edward Elgar, Cheltenham

Entorf H, Spengler H (2015) Crime, prosecutors, and the certainty of conviction. Eur J Law Econ 39(1):1–35

Farrington DP, Coid JW, Harnett LM, Joliffe D, Soteriou N, Turner RE, West DJ (2006) Criminal careers up to age 50 and life success up to age 48: new findings from the Cambridge Study in Delinquent Development, vol 299, Home office research study. Home Office, London

Garoupa N (2003) Behavioral economic analysis of crime: a critical review. Eur J Law Econ 15(1):5–15

Hallwood P, Miceli TJ (2014) Piracy, modern maritime. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York

Katz L, Levitt SD, Shustorovich E (2003) Prison conditions, capital punishment, and deterrence. Am Law Econ Rev 5(2):318–343
Ko H (2014) Behavioral law and economics. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York
Lando H (2006) Does wrongful conviction lower deterrence? J Leg Stud 35(2):327–337
Leroch MA (2014) Crime (expressive) and the Law. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York
Lochner L (2004) Education, work, and crime: a human capital approach. Int Econ Rev 45(3):811–843
Miceli TJ, Bucci C (2005) A simple theory of increasing penalties for repeat offenders. Rev Law Econ 1(1):71–80
Miles S, Pyne D (2014) The economics of scams. Presented at Canadian Economic Association, Conference, Vancouver
Miles S, Pyne D (2015) Deterring repeat offenders with escalating penalty schedules: a Bayesian approach. Econ Gov 16(3):229–250
Neilson WS, Winter H (1997) On criminals' risk attitudes. Econ Lett 55(1):97–102
Polinsky AM, Shavell S (2007) Public enforcement of law. In: Polinsky AM, Shavell S (eds) Handbook of law and economics. Elsevier, Amsterdam
Prescott JJ (2015) Criminal sanctions and deterrence. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York
Pyne D (2004) Can making it harder to convict criminals ever reduce crime? Eur J Law Econ 18(2): 191–201
Pyne D (2010) When is it efficient to treat juvenile offenders more leniently than adult offenders? Econ Gov 11(4):351–371
Pyne D (2012) Deterrence: increased enforcement versus harsher penalties. Econ Lett 117(3):561–562
Rizzolli M, Stanca L (2012) Judicial errors and crime deterrence: theory and experimental evidence. J Law Econ 55(2):311–338
Schneider F, Meierrieks D (2014) Terrorism. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York
Stigler GJ (1970) The optimum enforcement of laws. J Polit Econ 78(3):526–536
van Winden F, Ash E (2012) On the behavioral economics of crime. Rev Law Econ 8(1):181–213
Vannini M, Detotto C, McCannon B (2015) Ransom kidnapping. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York
Vollaard B and J Hamed J (2012) Why the police have an effect on violent crime after all: evidence from the British Crime Survey. Journal of Law and Economics 55(4):901–924
Wright R, Tekin E, Topalli V, McClellan C, Dickinson T, Rosenfeld R (2014) Less cash, less crime: evidence from the electronic benefit transfer program, vol 19996. National Bureau of Economic Research, Cambridge, MA

Crime, Unemployment and

Kangoh Lee
Department of Economics, San Diego State University, San Diego, CA, USA

C

Definition

Crime is one of the most important social issues citizens are concerned about. As such, it has received a good deal of attention from policymakers and scholars across fields such as criminology, economics, law, psychology, and sociology. Crime is partly motivated by economic conditions, and a large body of research has studied the relationship between economic factors, particularly unemployment, and crime for more than 50 years. Available evidence shows that the relationship is ambiguous, and it is an actively researched topic.

Introduction

Crime was first analyzed by economists in the late 1960s and early 1970s (e.g., Becker 1968; Ehrlich 1973). The analysis is based on the cost and benefit of crime. The cost to an individual who commits property crime is the lost legitimate incomes and the penalties when crime is caught, and the stolen properties and incomes constitute the benefit (see Anderson 2017 for a general discussion of the cost). The individual then compares the cost and the benefit when deciding to commit crime.

Applying the cost-benefit analysis of crime to the relationship between unemployment and crime, an increase in unemployment during economic downturns should increase crime, as the increase in unemployment decreases the cost of crime, namely, the lost legitimate income. However, this seemingly obvious and intuitive prediction has not been empirically supported. Rather, empirical findings have been mixed, ranging from positive effects of unemployment on crime to no stable or statistically significant effect and to negative effects. The relationship between

unemployment and crime is thus intriguing and one of the most controversial topics in law and economics. At the same time, this discrepancy between economic theory and empirical findings indicates that the unemployment-crime nexus is more complicated than the standard economic theory of crime predicts and calls for more research.

Effects of Unemployment on Crime

Overview

Fleisher (1963) is the first study to examine the effect of unemployment on crime. Based on data for property crimes from FBI Uniform Crime Reports over the period 1932–1961 across Boston, Chicago, and Cincinnati, he finds that the elasticity of the arrest rates with respect to the unemployment rate of young males is between 0.10 and 0.25. Since his work, a large number of research papers have empirically studied the relationship between labor market conditions and crime. These studies typically utilize panel data across jurisdictions such as counties or states during a certain period of time. They typically use instrumental-variable approaches to take into account the possible endogeneity of unemployment and to correct for simultaneity between unemployment and crime and controls for other covariates such as ages, incomes, police expenditures, and metropolitan areas (e.g., Raphael and Winter-Ebmer 2001; Gould et al. 2002; Lin 2008).

The literature on the topic has significantly grown, and a good number of papers have reviewed this growing literature (Long and Witte 1981; Freeman 1983, 1995; Chiricos 1987; Levitt 2004; Blumstein and Wallman 2006). These reviews show that the effects of unemployment on crime are positive, and sometimes statistically insignificant, and sometimes even negative. For instance, Freeman (1983) reviews 15 empirical studies and finds that unemployment has positive and significant effects on crime only in four studies. Chiricos (1987), by contrast, finds that unemployment has significant positive effects on crime in a majority of 63 studies he reviews.

These diverse findings show that the relationship between unemployment and crime is complicated. At the same time, these findings reflect the possibility that the relationship also hinges on the types of crimes and the characteristics of those individuals who commit crime. Crimes of passion such as murder and rape would depend less on economic conditions than crimes driven by economic incentives such as burglary and other types of property crimes. The youths have fewer economic opportunities than adults, and their motivation to commit crime may differ from adults'. Social customs and cultures that may govern the incentives to commit crime play a role in shaping the relationship. It is thus useful to examine the relationship between unemployment and crime separately for different types of crimes, for different groups of individuals, and for different countries.

Types of Crimes

There are different ways of categorizing crimes, but for the purpose of this entry, two types of crimes, violent crimes and property crimes, are relevant. According to the FBI Uniform Reporting Program, violent crimes are the offenses involving force or threat of force and consist of murder, rape, robbery, and aggravated assault. Property crimes do not involve force or threat of force against the victims and include burglary, larceny, motor vehicle theft, and arson.

The standard economics argument dictates that an increase in the unemployment rate in general should increase property crime but may not have much effect on violent crime. The reason is that those who commit property crime are interested in the properties of the victims rather than the victims themselves, and the economic conditions of the offenders such as their wages and employment status and those of the properties such as their market values affect the incentives to commit property crime. By contrast, violent crimes are the offenses against the victims, and the economic conditions of the offenders or the victims would not affect the incentives to commit violent crime in an important manner.

Recent empirical studies have on average confirmed the standard economics argument, but have

found that unemployment does not have uniform effects on property crime or violent crime (Raphael and Winter-Ebmer 2001; Gould et al. 2002; Lin 2008). In particular, an increase in unemployment is on average associated with a significant increase in property crime but has little effect on violent crime. However, an increase in the unemployment rates increases assault and robbery, but decreases murder and rape (Raphael and Winter-Ebmer 2001). In addition, the unemployment rate has a positive and significant effect on burglary and larceny, but a negative and significant effect on auto theft (Gould et al. 2002). Thus, unlike the standard economics argument, an increase in unemployment may affect some types of violent crimes in a significant manner and may decrease rather than increase some types of property crimes.

As these studies show, crimes driven by economic incentives, particularly burglary, appear to be strongly affected by unemployment, but the linkage between unemployment and violent crime is weak. If unemployment has any stable and expected effect on violent crime, it would affect robbery in a predictable manner, because robbery is a violent crime but close to property crime in its nature.

Different Countries

Diverse empirical findings may be attributed to different cultures and social norms, as the decision to commit crime may depend in part on cultures. This section discusses empirical findings across different countries to separate the effect of unemployment from the role of cultures and other social influences in crime. The literature mentioned above focuses on US data, and other countries and cross-country studies are considered below.

A number of studies have examined the link between unemployment and crime in Europe. In Germany, the effect of unemployment differs between before and after reunification. In particular, unemployment has little effect on crime in West Germany but has significant and positive effects on robbery and theft in reunified Germany (Entorf and Spengler 2000). In Sweden, unemployment has the expected effects on crime in the sense that unemployment has a positive and significant effect on property crime, but no significant effect on violent crime (Edmark 2005). In England and Wales, an increase in unemployment decreases fraud and increases drug and other crimes among ten police-recorded crimes (Wu and Wu 2012).

As for countries other than the USA and European countries, a positive and significant relationship between unemployment and crime has been found in New Zealand (Papps and Winkelmann 2000), in Malaysia (Tang 2011), and in some Latin American cities but not in other Latin American cities (Hojman 2004). As for multi-country studies, Wolpin (1980) explores three-country data, Japan, the UK, and the USA, and demonstrates that unemployment has a positive effect on crime, and Altindag (2012) analyzes a set of European cross-country data and shows that unemployment increases crime.

The above studies show that the unemployment-crime connection has been extensively investigated across countries, reflecting global interests in the topic. At the same time, judging from their findings, it appears that the relationship between unemployment and crime does not crucially depend on countries in the sense that unemployment largely tends to increase property crime.

Groups of Individuals

Youth crime has been an important subject of crime research in general, as youths are more likely to commit crime than adults and face different economic opportunities. A number of studies consider the effect of youth unemployment on youth crime, and the effect appears to differ from the standard results in the literature above. In particular, it has a long-run positive effect on fraud, homicide, and motor vehicle theft but not on other types of crimes in Australia (Narayan and Smyth 2004), and it increases burglary, theft, and robbery, but not violent crime in England and Wales (Carmichael and Ward 2001). In France, the unemployment rates of 15–24-year-olds increase almost all types of crimes, including violent crimes, but the unemployment rates of 25–49-year-olds decrease almost all crimes, again including violent crimes, although the effects of the

unemployment rates of the latter group are not statistically significant (Fougère et al. 2009).

A rare study, based on Florida county-level data, investigates the effects of unemployment on reconviction for black males and for white males separately (Wang et al. 2010). They find that black ex-prisoners released to areas with high black-male unemployment rates are more likely to commit violent crime, but no such effect is found for white counterparts. They also find that white ex-prisoners released to areas with high white-male manufacturing employment rates are less likely to commit violent crime, but no such effect is found for black counterparts. It is interesting to observe that unemployment rates or employment rates of the areas have no effect on the likelihood of ex-prisoners committing property crime.

The effects of youth unemployment on youth crime appear to differ in the sense that youth unemployment also affects youth violent crime at least in France and Australia. Many studies include the proportion of blacks in a county or region or state as a covariate in their regressions, but few studies consider blacks and whites separately, and the study by Wang et al. (2010) shows that significant differences exist between blacks and whites in terms of the effects of unemployment on recidivism. Another group of individuals of interest are females. Females are known to be less likely to commit crime than males, but no research appears to examine the relationship between female unemployment and female crime.

Moderating Factors

Many studies on the relationship between unemployment and crime control for other covariates, as noted above, but the relationship itself depends on moderating variables. For instance, the effect of unemployment on crime hinges on the apprehension rate (Lee 2016). In particular, an increase in the unemployment rate decreases the crime rate at low apprehension rates but increases the crime rate at high apprehension rates, as apprehension affects both the cost of crime and the benefit of crime.

In regression analysis, moderating variables are captured by interaction terms. While interaction terms have been rarely studied in the literature, Baron (2008) includes an interaction term between youth unemployment and monetary dissatisfaction in the youth violent-crime regression and an interaction term between youth unemployment and job search activities in the youth drug-crime regression. The regression results show that an increase in youth unemployment tends to increase youth violent crimes when youths are more dissatisfied with their monetary situations, and it tends to decrease youth drug crimes when they search for jobs more regularly.

Other Labor Market Conditions and Crime

Among labor market opportunities that may affect crime, unemployment has been the most important determinant of crime. However, other labor market conditions also play an important role in crime. An increase in wages of noncollege-educated or college-educated men reduces crimes broadly, including both property crime and violent crime in the USA (Gould et al. 2002). A decrease in the wages of low-wage workers substantially increases all types of property crime in England and Wales (Machin and Meghir 2004).

Income inequality is another factor that affects crime. The way income inequality affects property crime and violent crime differs significantly from the way unemployment does. More inequality significantly increases violent crime but has no effect on property crime in the USA (Kelly 2000). This finding stands in sharp contrast with the result that unemployment has a positive effect on property crime, but little effect on violent crime. One interpretation of the contrast is that property crime is influenced by economic incentives while violent crime is driven by strain and social situations. The income inequality may affect crime differently in a time-series analysis. In particular, an increase in inequality increases crime rates in the cross-section analysis but decreases crime rates in the time-series analysis (Brush 2007). In a study that relates inequality and unemployment to crime in Latin American cities, Hojman (2004) finds that more inequality significantly increases

crime in Greater Buenos Aires. It is difficult to directly compare Kelly (2000) with the last two studies, as they do not distinguish property crime from violent crime, but income inequality appears to influence crime in general although the inequality-crime nexus has not been extensively studied.

Conclusions

Crime causes private costs to victims and criminals and social costs as well, and policymakers have attempted to reduce crime by spending resources. To the extent that there is a connection between crime and economic incentives, any crime policy has to examine such economic incentives. This entry has focused on unemployment as the key part of economic incentives. Despite a large volume of empirical research on the topic, the literature has not reached a consensus on the effects of unemployment on crime even though the standard economics argument dictates that unemployment must have a positive effect on crime. Rather, the effects depend crucially on the types of crimes and on the characteristics of individuals as well. The literature is still growing and expected to attract attention from scholars and policymakers.

While unemployment is important, other labor market conditions such as wage levels and wage inequality appear to influence the decisions to commit crime. Nevertheless, little research on the wage-crime link or the inequality-crime link has been conducted, and more research on the issue is warranted. In addition, many interesting aspects of crime have been omitted in this entry, but other entries discuss them (e.g., Baker 2017; Leroch 2017; Prescott 2017; Pyne 2017).

References

Altindag DT (2012) Crime and unemployment: evidence from Europe. Int Rev Law Econ 32:145–157

Anderson D (2017) Cost of crime. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York. forthcoming

Baker M (2017) Crime (organized) and the law. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York. forthcoming

Baron SW (2008) Street youth, unemployment, and crime: is it that simple? Using general strain theory to untangle the relationship 1. Can J Criminol Crim Justice 50:399–434

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 73:169–217

Blumstein A, Wallman J (2006) The crime drop and beyond. Ann Rev Law Soc Sci 2:125–146

Brush J (2007) Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties. Econ Lett 96:264–268

Carmichael F, Ward R (2001) Male unemployment and crime in England and Wales. Econ Lett 73:111–115

Chiricos T (1987) Rates of crime and unemployment: an analysis of aggregate research evidence. Soc Probl 34:187–212

Edmark K (2005) Unemployment and crime: is there a connection? Scand J Econ 107:353–373

Ehrlich I (1973) Participation in illegitimate activities: a theoretical and empirical investigation. J Polit Econ 81:521–565

Entorf H, Spengler H (2000) Socioeconomic and economic factors of crime in Germany: evidence from panel data of the Germany states. Int Rev Law Econ 20:75–106

Fleisher BM (1963) The effect of unemployment on juvenile delinquency. J Polit Econ 71:543–555

Fougère D, Kramarz F, Pouget J (2009) Youth unemployment and crime in France. J Eur Econ Assoc 7:909–938

Freeman RB (1983) Crime and unemployment. In: Wilson JQ (ed) Crime and public policy. Institute for Contemporary Studies Press, San Francisco, pp 89–106

Freeman RB (1995) The labor market. In: Wilson J, Petersilia J (eds) Crime: public policies for crime control. Institute for Contemporary Studies Press, San Francisco, pp 171–192

Gould E, Weinberg B, Mustard D (2002) Crime rates and local labor market opportunities in the United States: 1979–1997. Rev Econ Stat 84:45–61

Hojman DE (2004) Inequality, unemployment and crime in Latin American cities. Crime Law Soc Chang 41:33–51

Kelly M (2000) Inequality and crime. Rev Econ Stat 82:530–539

Lee K (2016) Unemployment and crime: the role of apprehension. Eur J Law Econ. Forthcoming. https://doi.org/10.1007/s10657-016-9526-3

Leroch M (2017) Crime (expressive) and the law. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York. forthcoming

Levitt SD (2004) Understanding why crime fell in the 1990s: four factors that explain the decline and six that do not. J Econ Perspect 18:163–190

Lin M-J (2008) Does unemployment increase crime? Evidence from U.S. data 1974–2000. J Hum Resour 43:413–436

Long SK, Witte AD (1981) Current economic trends: implications for crime and criminal justice. In: Wright KD (ed) Crime and criminal justice in a declining

economy. Oelgeschlager, Gunn and Hain, Cambridge, MA, pp 69–143
Machin S, Meghir C (2004) Crime and economic incentives. J Hum Resour 39:958–979
Narayan PK, Smyth R (2004) Crime rates, male youth unemployment and real income in Australia: evidence from Granger causality tests. Appl Econ 36:2079–2095
Papps K, Winkelmann R (2000) Unemployment and crime: new evidence for an old question. N Z Econ Pap 34:53–71
Prescott JJ (2017) Criminal sanction and deterrence. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York. forthcoming
Pyne D (2017) Crime (incentive to). In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer, New York. forthcoming
Raphael S, Winter-Ebmer R (2001) Identifying the effect of unemployment on crime. J Law Econ 44:259–283
Tang CF (2011) An exploration of dynamic relationship between tourist arrivals, inflation, unemployment and crime rates in Malaysia. Int J Soc Econ 38:50–69
Wang X, Mears DP, Bales WD (2010) Race-specific employment contexts and recidivism. Criminology 48:1171–1211
Wolpin KI (1980) A time series-cross section analysis of international variation in crime and punishment. Rev Econ Stat 62:417–423
Wu D, Wu Z (2012) Crime, inequality and unemployment in England and Wales. Appl Econ 44:3765–3775

Crime: Economics of, Different Paradigms

Sven Grüner and Norbert Hirschauer
Agribusiness Management,
Institute of Agricultural and Nutritional Sciences,
Martin Luther University Halle-Wittenberg,
Halle (Saale), Germany

Definition

Economics of crime describes the attempt to explain rule-breaking behaviors based on the assumption that people make purposive choices under conditions of scarcity.

Economics of Crime: Different Paradigms

Economics of crime can be broadly defined as the attempt to explain rule-breaking behaviors based on the assumption that people make purposive choices under conditions of scarcity. People's choice sets regularly contain both permissible and illicit choices. Illicit choices do not always constitute criminal acts in terms of violations of the criminal law. However, throughout this entry we use crime, illegal behavior, illicit choice, noncompliance, offense, and rule breaking as interchangeable terms for the infringement of formal (codified) rules. **Economics of crime** includes the analysis of **economic crime** and white-collar crime such as corruption, patent infringements, or the violation of industrial safety laws, but it explicitly applies economic approaches to the study of illicit behaviors such as traffic offenses and vandalism, which are often considered being beyond the realm of economics.

Economic approaches share the core understanding that human behaviors are the result of purposive choices – either fully rational (e.g., von Neumann and Morgenstern 1944) or 'intendedly rational' (Simon 1955: 114). However, there is no uniform conception of economic man. Instead, differing models – from purely materialistic and self-interested **rational choice** to multigoal and **bounded-rational choice** – are used to study behaviors in different contexts. This entry systematically discusses competing conceptions (paradigms) of illicit choice.

Beccaria, Bentham, Becker

Gary S. Becker (1968) is one of the most prominent, but by far not the first one to study both crime (self-interested private choice) and crime prevention (welfare-oriented public choice) from an essentially economic (utilitarian) perspective. There are many precursors of Becker in modern history who applied the logic of utilitarian calculus to the study of crime and its **prevention through deterrence**. The most noteworthy one is Beccaria whose famous treatise *Dei delitti e delle pene* (*On Crime and punishment*; 1764/1995) combined utilitarian reasoning with the Enlightenment call for the rule of law and the plea against excessive and disproportionate punishment. Drawing on Hobbes' social contract

theory (1651/2010), Beccaria justified state-imposed punishment as the necessary means to enforce the social contract as codified in the law. Beccaria's ideas on punishment heavily influenced Bentham (1789). Beccaria and Bentham agreed on three fundamentals: (1) the need for legal formalism to justify the justice system, (2) the rejection of capital punishment due to its brutalizing effects on society, and (3) the endorsement of a preventive instead of a retributive rationale in sentencing that limits punishment (Harcourt 2014). Antedating modern-day understanding of deterrence as being both specific and general, Beccaria (1764/1995: 31) detailed that the only purpose of punishment is to "prevent the offender from doing fresh harm to his fellows and to deter others from doing likewise." He also noted that, from a utilitarian point of view, there is a moral-philosophical limitation to punishment: "If a punishment is to serve its purpose, it is enough that the harm of punishment should outweigh the good which the criminal can derive from the crime, [. . .]. Anything more than this is superfluous and, therefore, tyrannous" (Beccaria 1764/1995: 64).

The general idea of **Beccarian deterrence** is that presumptive criminals will respond to the **severity** ("size") of punishment, its **certainty** (probability), and its **celerity** (swiftness). Beccaria and Bentham strongly advocated mild but highly probable and swift punishment. They argued that certainty and swiftness are necessary to make potential criminals realize that crime is closely associated with punishment. In their understanding, the fear of a more severe but less likely and delayed punishment, which inevitably goes hand in hand with the hope of not being captured at all, will deter crime less effectively. They furthermore noted that overly severe penalization might harden criminals and provoke follow-up crimes to avoid capture and hard punishment.

Neoclassical accounts of crime based on twentieth-century microeconomic theory, such as those of the Chicago School of economics and notably Becker (1968), took up the Beccarian idea of utilitarian calculus and deterrence. Becker modeled individuals as materialistic, self-interested, and rational decision makers who maximize their utility. In line with the **neoclassical paradigm**, Becker supposed people's preferences to be context-independent ("exogenous") and stable. He assumed that individuals calculate and weigh the material benefits (utility) and costs (disutility) of rule breaking and rule abidance without moral considerations. Based on this assumption, Becker arrived at normative regulatory conclusions through the **externality** argument. He contended that the **expectation value of the sanction** ("expected sanction") – as resulting from its size and probability – should be set at a level that makes would-be offenders internalize the negative externalities (harms) they would inflict on the victims. Using the language of price theory, Becker claimed that, to decrease the "supply" of crime, regulators need to engage in monitoring and sanctioning activities (enforcement) to reduce the excessive "price" that criminals could otherwise obtain from illegal activity. Regulators should undertake these law enforcement activities up to the level where their marginal social costs equal the marginal social costs of criminal behavior. This implies tolerating a certain amount of crime because, at some point, the costs of more law enforcement would exceed its benefits.

Becker's normative claim that sanctioning should make offenders internalize the harms inflicted on others was challenged from within the neoclassical camp. Posner (2014) claimed that the expected sanction should prevent crime altogether, by fixing it at a level that (marginally) surpasses the offender's illicit profit. According to Becker's *internalization approach*, one should tolerate "**efficient**" **crimes** (the breaking of inefficient rules) that lead to a more efficient allocation of resources (Harel 2014). A polluter whose illicit profit exceeds the environmental damage should not be deterred, for example. In contrast to Becker but in line with Beccaria, Posner's *prevention approach* implies that any crime (injury of a legally protected interest) should be prevented by turning it into an inferior option for the agent deliberating it.

Becker focused on "**taxing**" **rule breaking** as *the* means to reduce its "price" relative to the one

attached to legal activity. However, as noted by himself, the attractiveness of rule breaking can also be reduced by "**subsidizing**" **rule abidance** and increasing the opportunity costs of crime – for example, by providing education and legal income opportunities to destitute people who are at risk of becoming offenders. The deterrence argument leaves open the empirical question whether money spent on monitoring and sanctioning (taxing illegal behavior) deters crime more efficiently than money spent for subsidizing legal behavior. Nonetheless, Becker and many other advocates of "negative" deterrence seem to focus from the outset on harsh penalization.

While Becker (1968) praised Beccaria as the first one to have used economic calculus in the study of crime, their ideas differ markedly: first, Becker favored capital punishment. Second, he did not elaborate on the effects of the celerity of capture and punishment. Third, he did not believe that *low-size-high-probability sanctioning* (**Beccarian deterrence**) deters crime more effectively that than *high-size-low-probability sanctioning*, as proposed by himself (**Beckerian deterrence**). Based on expected utility theory, he argued that Beccarian deterrence would only work better if people were risk preferring (which is not plausible). In contrast, increasing the size of punishment and decreasing its probability, while maintaining its expectation value, will reduce the utility of illegal behavior and thus deter crime more effectively if people are risk averse (which is plausible). Becker claimed that his sanctioning regime would be superior even if people were risk neutral because it would deter crime as effectively as Beccarian deterrence but allow regulators to economize on enforcement costs. Kolm (1973: 266) pointedly highlighted the weak spot of the Beckerian conclusion as "*hang offenders with zero probability.*"

The legal doctrine regarding punishment: Despite its label, *economics of crime* is commonly seen as encompassing the study of criminal *and* noncriminal offenses. This loose terminological usage would not be acceptable for legal scholars.

Criminal lawyers will note that the penal code differs substantially from other codes of laws and that its provisions must be limited to reflect society's most fundamental moral sentiments of right and wrong. Inversely, it is to provide a stigma capable of causing social costs to criminals and shaping the norms of people. That is, less forceful legal institutions such as administrative and tort law should be used to outlaw less serious misconducts. Otherwise, the reciprocal links between the criminal law and society may erode which, in turn, may cause perverse behavioral effects and the degradation of the criminal law system.

Legal doctrine (cf., Radbruch 2011) must reject Becker's normative conclusions for three reasons. *First*, Becker's internalization approach transfers the liability rationale of tort law to criminal law. Like administrative law, tort law does not provide a strong stigma. Hence, many ordinary (noncriminal) people find it acceptable to occasionally break civil or administrative law provisions (minor torts or violations of building law, traffic law, etc.). Becker suggested that even crime be accepted if it facilitates an efficient allocation of resources. The criminal law, however, must stigmatize crime – be it "efficient" or not – more than other transgressions because crime not only reallocates resources but also violates the inalienable individual rights of a person and the universal rights of society as a whole. *Second*, legal scholars will not endorse the all-dominant role of the expectation value of the sanction, and they will reject the suggestion to adjust the size of a sanction contingent on its probability of being imposed. The justice system must not only produce deterrence but also be proportionate and fair to the individual person (proportionality principle). The rule of law requires a transparent uniformity of the law aimed at meeting equal crimes with equal punishment. This precludes subjecting individuals

committing minor transgressions such as parking offenses to extreme punishment ("hanging") on the sole ground that detection probability is low. *Third*, practically minded legal scholars will add that it is not possible to differentiate sanctions contingent on specific enforcement contexts. Imagine a city council that reduces the number of police officers for budgetary reasons. In Becker's view, this would require adjusting the sanction catalog in this city and punishing offenders, who are caught despite slack enforcement, more severely. How should the lawmaker learn about each city's specific circumstances? What about law adjustment costs? Moreover, how should presumptive offenders learn about the sanctions in place?

It seems that the normative recommendations of conventional *economics of crime* are often not feasible from a legal doctrine point of view. This is because conventional *economics of crime* pays notoriously little attention to legal requirements that limit the set of deterrence strategies available to regulators. Following its recommendations would more often than not violate fundamental legal principles and compromise the rule of law.

Neoclassical approaches to deterrence are not naïve. They do not deny that individual behaviors are frequently incompatible with the axioms and predictions of rational choice. They contend, however, that deviations from the standard model are minor or at least symmetrical (Korobkin and Ulen 2000). In their view, it is therefore adequate to model people as rational and self-interested utility maximizers even if there is not a single individual in the world who is a fully rational egoist. For illustration sake, let us look at the "supply" of a criminal act such as theft. If people's evaluations were randomly distributed around a mean defined by a risk averse rational egoist, the utility obtained from theft (and thus its "supply") could be predicted to decrease if regulators, while maintaining the expectation value of the sanction, compensated low risks of punishment with high sanction levels. However, this will not hold if the regulatees' *perception* of the detection risk is *asymmetric* because they underrate or even neglect small probabilities. In this case, Beckerian (high-size-low-probability) sanctioning will prevent less crime than Beccarian (low-size-high-probability) sanctioning. If people's judgments are affected by such **systematic biases** (either generally or in certain social groups), understanding these biases will enable regulators to better predict and eventually influence people's behavior.

In Beccaria and Bentham's view, highly certain and swift punishment is needed to make people fully *perceive* the association of crime with punishment. These early pioneers seem to have anticipated cognitive features that behavioral economists today would classify as forms of bounded rationality, namely, "**nonlinear weighting of probabilities**" and "**hyperbolic discounting**." Beccaria and Bentham apparently realized that people might underrate or neglect small probabilities and be biased towards the present. When both effects combine, people will attribute a high value to the immediate benefits of crime and a very low or no value to its high but uncertain future costs. This seems a plausible explanation of many rule breaking instances including illicit drug use with its mainly self-inflicted future punishment of bad health or even premature death. Nonetheless, Becker disregarded bounded-rational judgments and was convinced that one can dispense with all other theories including those of "psychological inadequacies." He believed *his* theory of crime and emphasis on harsh punishment to be an improvement on not only Beccaria but all crime scholars who do not exclusively rely on the neoclassical paradigm. Due to frustrating experiences with harsh punishment, many scholars today think that Beccaria was right to advocate mild but certain and swift punishment. Nevertheless, Becker and his followers, backed up by the electorate's liking for harsh punishment, seem to have had an influence that partly reverted crime policies, especially in the USA, back to premodern penalization

regimes with capital punishment, long prison sentences, and mass incarceration. Even though heavily criticized on methodical grounds, a time-series analysis by Ehrlich (1975), claiming that capital punishment substantially reduced capital crimes in the USA, fueled this development. Using more adequate econometric approaches, more recent studies did not find that capital punishment deters capital crime (e.g., Zimmerman 2004; Kovandzic et al. 2009).

Behavioral Economics of Crime

There is considerable evidence from experimental and observational studies that, for various reasons, people's behaviors deviate substantially *and* systematically from rational choice predictions (Thaler 2016). This is why regulatory strategies exclusively based on the neoclassical paradigm, which is still mainstream in law and economics, are a socially inefficient way to prevent illicit behaviors in many contexts. An educated guess would be that they could work for tax offenses and securities fraud but less so for street and property crimes or drug offenses.

A **behavioral-economics-of-crime** analysis that provides a reconstructing understanding of perpetrators' judgments is needed to close the gap between rational choice predictions and actual behavior. Instead of relying on rational choice and objective facts, **reconstructing understanding** implies trying to understand people's options of choice and calculi according to their subjective perceptions and evaluations. As Rubinstein pointed out (1991: 910), "these [perceptions] need not necessarily represent the physical rules of the world." As people are heterogeneous in both their goals and judgments, making reliable predictions and identifying adequate enforcements strategies is a challenging task. It is also a promising task because many laws address specific groups whose members share behavioral regularities. Such regularities can be identified and considered when specifying contextual enforcement strategies. Adequate strategies will substantially differ, for example, between fields such as securities legislation, traffic law, or drug law. A promising regulatory potential termed "**nudge**" by Thaler and Sunstein (2008) arises from the fact that people's behaviors often depend on how contexts are described or "framed" (Tversky and Kahneman 1981). Finally, people's goals and judgments change and can be shaped (to a certain degree) over time.

Table 1 itemizes a selection of deviations from rational choice that seem especially relevant for the study of illegal behaviors. The large number of "anomalies" – as they would be labeled from the neoclassical point of view – forces us to limit this entry to a brief and focused description of those deviations that seem to be most relevant for understanding compliance decisions. Further descriptions regarding behavioral economics applications with respect to noncompliance can be found in Dhami (2016), Thaler (2016), or Korobkin and Ulen (2000), for example.

Multiple Goals, Social Norms, and Bounded Self-interest

People's utility does not exclusively hinge on material outcomes including leisure and

Crime: Economics of, Different Paradigms, Table 1 Deviations from rational choice relevant for the study of crime

Multiple goals, social norms, and bounded self-interest	Bounded-rational judgments		
	Cognitive biases	**Heuristics**	**Bounded self-control**
Striving for conformity with external social expectations	Non-linear weighting of probabilities	Habits	Hyperbolic discounting
Striving for consistency with internalized values and norms	Overconfidence bias	Adhering to defaults	Emotions
	Loss aversion	Mental accounting	Limited attention

conveniences. In contrast, people have complex **multidimensional goal systems** that are influenced by the apparently innate human urge to repay both good and bad experiences in kind (**reciprocity**). Besides the self-interested pursuit of material benefits, people particularly strive for two types of intangible outcomes: (1) conformity with external social expectations and (2) consistency with internalized values and norms. These intangible goals can limit people's self-interest (**bounded self-interest**). Different labels have been attached to these two behavioral drivers. Psychologists often speak of an extrinsic as opposed to an intrinsic motivation to comply. Coming from an applied management research background, Nielsen and Parker (2012) use the terms "social preferences" as opposed to "normative preferences." Adopting an institutional economics perspective on rule-governed social life in general, Ostrom (2005) speaks of external as opposed to internal "*delta* parameters" to indicate that social norms have to be considered as behavioral drivers in *addition* to material incentives.

Criminology with its primary focus on the law and law enforcement provides still another terminology and talks of "protective factors" and "risk factors" (Hirschauer and Scheerer 2014). **Protective factors** are bonds to social norms that discourage illicit choices by causing nonmaterial costs for rule-breaking and nonmaterial benefits for rule abidance. Protective factors can be seen as mechanisms that impose "intangible taxes" on illegal acts and provides "intangible subsidies" for legal acts. These taxes and subsidies arise from external social control (social disapproval/ostracism vs. social approval) as well as from internalized values (self-disrespect/guilt vs. self-esteem). Protective factors can be related to the concept of **resilience**. Instead of seeing people as being either law-abiding or criminal, this implies understanding people as being more or less resilient to moral hazards in that they have a positive but limited willingness-to-pay for rule abidance in exchange for social approval and a feeling of integrity.

In contrast, **risk factors** arise in deviant subcultures in which social norms are not in line with those of the law-approving "conventional" society. Examples are youth gangs, organized crime communities, or groups hostile to state authority due to extreme religious or ideological beliefs. Like protective factors, risk factors can arise from external and internal sources. They act as "intangible taxes" for legal behaviors (disrespect by deviant peers and conflict with the deviant internal self) and as "intangible subsidies" for illegal behaviors (respect by deviant peers and affirmation of the deviant internal self). Altruistic rule-breaking and **reactance** (Miron and Brehm 2006), for example, are internal risk factors resulting from a deviant self. Even though frequently neglected, risk factors and protective factors are often decisive for compliance decisions. Group-specific social pressures and rewards to break the law, in conjunction with weak conventional norms against illegal behavior or outright deviant selves, for example, are likely to be more relevant causes of juvenile delinquency, such as illegal road races or vandalism, than illicit profits. The interplay of protective factors and risk factors is also crucial for understanding how individual employees behave within deviant corporate cultures (multiple-selves problem).

Crime prevention is an intentional manipulation of factors that determine people's behavior with regard to the law. Three ideal-type strategies are distinguished (Picciotto 2002): **incapacitation** (e.g., through license withdrawal or jailing) is aimed at reducing would-be perpetrators' opportunities to commit offenses. **Deterrence** (e.g., through monitoring and sanctioning) is to reduce the economic temptations (incentives) for rule breaking. **Accommodation** (e.g., through persuasion and advice) is to bolster people's propensity to comply by strengthening the social norms that back up the rules (protective factors). For example, informing defaulting taxpayers that the majority of people in the same town have already paid their tax debts was found to have a positive effect on tax compliance (The Behavioural Insights Team 2011).

Identifying adequate prevention strategies requires a normative analysis based on conditional forecasting. In this forecasting exercise, the behavioral effects of economic incentives and norm-based "taxes" and "subsidies" cannot

be considered in isolation. Reducing the economic attractiveness of rule breaking through tight controls and harsh sanctions may reinforce social norms in some instances but weaken them in others. The label "**crowding in**" describes interventions that simultaneously provide the "right" incentives and strengthen desirable social norms. Evidence from fields as far apart as industrial safety and tax legislation indicates that successful strategies manage to crowd in (Braithwaite 2009). They avoid the dysfunctional effects of oppressive control and harsh sanctions (e.g., reactance) as well as the negative effects of overly lenient approaches (blurring of standards). In contrast, "**crowding out**" describes interventions that involuntarily weaken desirable social norms in the attempt to reduce the economic temptations for rule breaking (Frey 1997). An illustrative example comes from a field experiment in daycare centers (Gneezy and Rustichini 2000). After a small fine was introduced for collecting children late, the frequency of late pickups increased. The interpretation by Sandel (2012) is that under the old regime parents tried hard to be in time because they were ashamed to cause inconvenience for the staff. After the introduction of the fine, which doubtlessly increased the economic costs of coming late, parents felt no shame anymore but came to see late pickups simply as an extra service for which they paid a "fee."

The intended primary effect of criminal sanctioning is often not economic deterrence. Instead, one hopes that the social stigma provided by the penal law, such as banning the beating of children, will reinforce desirable social norms. Including offenses into the penal code may backfire if the values of the lawmaker and of those subjected to the law do not correspond. Criminalizing activities that many people do not consider "real" wrongs may weaken the criminal law's general capacity of cultivating social norms. What is more, it might provoke **perverse effects** beyond those caused by the weakening of crime-inhibiting social norms. For example, criminal sanctions on the use of soft drugs, while increasing its cost, might actively encourage substance abuse if users develop individual reactance or even group norms that favor defying what they consider illegitimate and oppressive government intervention.

With a focus on corporate governance and recidivism, Braithwaite (2002) emphasized that a transparent progression of regulatory responses – accommodation first, deterrence second, incapacitation third – is needed to fully exploit the steering potential of regulatory enforcement. Using the terms "**enforcement pyramid**" and "**responsive regulation**," he claimed that regulators should meet noncompliance always with a clear disapproval and the request to remedy the wrong. They should use increasingly severe deterrent measures (warning letters, tightening controls, increasing fines) if offenders continue to break the rules. As a last resort, recidivists should be subjected to increasing levels of incapacitation (license suspension, license revocation, jailing). Braithwaite's key message is that regulators should aim to reintegrate offenders into rule-abiding society by using *transparent* and *graduated* response measures contingent on the degree of bad/good conduct. While regulators should always start softly, the concept of responsive regulation, which is inherently aimed at crowding in, is not leniency. On the contrary, it assumes that the harsher the available ultimate sanctions, the more likely regulators will achieve compliance through persuasion. The recommendation to regulators would therefore be to "speak softly, while carrying very big sticks" (Ayres and Braithwaite 1992: 40).

Bounded Rational Judgments

Besides social norms and bounded self-interest, bounded-rational choice needs to be considered when trying to understand and steer people's behavior. The term "**bounded rationality**" was coined by Simon (1957) to describe that individuals have limited access to decision-relevant information and that they have limited abilities to process that information. Consequently, choices are often biased and may depend on the presentation of the decision problem (framings) even if the framings have no effect whatsoever on outcomes. To reduce complexity and cope with their bounded-rationality, people often use simplifying decision rules or "heuristics"

(Gigerenzer and Todd 1999). **Cognitive biases**, **heuristics**, and **bounded self-control** can lead to choices that are contrary to rational choice expectations according to which fully informed people judiciously weigh the costs and benefits associated with their choices.

Nonlinear weighting of probabilities: Expected utility theory assumes linearity in the weighting of probabilities. In contrast, Kahneman and Tversky (1979) claimed, on the one hand, that people frequently underweight high probabilities and overweight low probabilities. On the other, they noted that people's perception is reversed near the end points in that very high probabilities are perceived as certain, whereas very low probabilities are underrated or even ignored. In the context of crime, the latter misconception is sometimes fueled by "optimism bias" (Weinstein and Klein 1996), which describes a person's belief to be less prone to risks and negative events than others. An important source of such misconceptions are premature generalizations from personal experiences that are easily mentally available and thus salient ("availability bias"). Illicit behaviors associated with low sanctioning probabilities, such as traffic offenses or free-riding public transport without a ticket, are illustrative examples. They may become business as usual for offenders who repeatedly experience no controls. The ostensible remedy of making the risk of being caught more salient by disclosing the number of caught offenders may produce perverse effects, however. For example, while such a piece of information clearly points out to free-riders that the probability of being caught is *not* zero, nondelinquent individuals might adjust a previously overrated detection probability downwards which, in turn, might produce some new free-riders. What is more, obtaining the knowledge that free-riding is a common phenomenon might erode the social norm that backs up the formal obligation to purchase a ticket. In this context, Popitz (1968) spoke of the "preventive effect of ignorance."

Overconfidence bias: Offenders are not average representatives of the population. At least a certain subset of wrongdoers may overestimate their capabilities and believe that they are more apt and clever to commit crimes and avoid punishment than others. What is more, they may also have the disposition to interpret even quite unambiguous information in ways that fall into place with their preconceived notions ("self-serving bias"). Overconfidence and self-serving bias can explain "silly" crimes for which no explanation in terms of the criminal's self-interest can be found. Deterrence strategies that target presumptive offenders who exhibit overconfidence and/or self-serving bias need to strengthen the salience of controls and sanctions. This may require increasing controls and sanctions beyond the level that would be needed to deter unbiased offenders.

Loss aversion: People often do not think in terms of final states but rather in terms of changes that they perceive as gains or losses with respect to a reference point (Markowitz 1952). Kahneman and Tversky (1979) claimed that people experience losses about twice as strongly as gains. The perception of a change as either a gain or a loss depends on contexts/framings and can affect compliance. Imagine two tax regimes. In regime 1, a taxpayer has to pay 1,000 in income tax per month. In regime 2, the total tax debt of 12,000 is payable after the end of the tax year. In both regimes, tax dodgers run a 25% risk of being detected and fined to pay a penalty of 48,000 in addition to the 12,000 in due taxes. Rational choice theory would predict that taxes are paid lawfully in both regimes because the *expected* reduction in income amounts to 15,000 ($= 0.25 \cdot 60{,}000$) in the case of tax evasion, but only to 12,000 in the case of lawful tax payment. In addition, tax evasion would increase income *risk*. Things are different if the taxable person exhibits loss aversion. In regime 1, such a taxpayer is likely to perceive the income level *after* regular monthly tax payments as reference point and experience the taxes as nonrealized income gains (opportunity costs). In regime 2, the taxable person is likely to get used to the income level *before* paying taxes and correspondingly adjust the reference point upwards. From the increased reference point, paying the tax debt of 12,000 will be perceived as a certain loss (out-of-pocket costs) that, despite being identical, is experienced as a heavier burden than the taxes in regime 1.

Depending on the strength of this effect, the taxable person in regime 2 will evade taxes and exchange the certain loss for the uncertain chance to escape tax payments. From a rational choice point of view, this implies behaving like a risk seeker. Loss aversion can be related to status quo bias and, more particularly, the endowment effect. The latter describes the fact that people find it generally hard to part with goods and wealth they already enjoy and (believe to) legitimately own.

Habits: Depending on the familiarity with and the relevance of choices, individuals either rely on habitual behaviors or consciously resort to problem-solving procedures (Katona 1953). Each day, people need to make so many choices that they must be made fast. This is why most choices are not the result of analytical procedures. Instead, people cannot help but make most choices habitually and spontaneously. This has been labeled "thinking fast" by Kahneman (2011). Having a habit means making the same choices as in the past when dealing with recurring or similar decision problems. Habits are arguably the most important heuristic without which we would not be able to survive. Habitual behaviors can be important for understanding some types of illegal behaviors. Depending on the individual's life course, offenses such as speeding or even evading taxes may have become subconscious habits irrespective of whether they provide noteworthy benefits for the offender or not. Habits in conjunction with self-serving bias may pose serious obstacles to effective deterrence, which is inherently based on informing people about controls and sanctions. This is because people are prone to dissolve cognitive dissonances by aligning their interpretation of information to their deep-set habitual behaviors (self-manipulation of beliefs) instead of aligning their behaviors to the information.

Adhering to defaults: Default options seem to exercise a stronger influence on people's behavior than what we would expect when considering the usual small effort needed to opt against the default. An illustrative example is people's acceptance to become organ donors (Gigerenzer 2010). The willingness to donate organs is considerably lower in countries that use "no organ donation" as default compared to countries that apply the nudge device of making "organ donation" the default option. A plausible explanation is that individuals orient themselves to what they perceive as the socially desired standard. People's preference for the as-is state (the default) can also be understood as a habit or a special form of the "status quo bias."

Mental accounting: People occasionally classify their choices into categories or "accounts" (health account, environment account, etc.) within which they offset their choices against each other (Thaler 1999). Imagine dog owners who mentally enter "paying dog tax" and "cleaning up when walking the dog" to the same account. After the introduction of a dog tax, they might feel that they can offset their cleaning up against their paying the dog tax. The important message is that introducing new regulations can have negative side effects that are easy to overlook. The effect produced by mental accounting looks like crowding out at first view. The difference is that mental accounting describes how the levels of *different* activities, which people see as belonging to the same "account," affect each other. Crowding out, in contrast, looks at *one* activity and describes that incentives might involuntarily weaken favorable social norms.

Hyperbolic discounting: The conception of rational choice includes the concept of time preferences that assumes that people discount future outcomes by using their individual rate of time preference. They presumably do so, however, by using constant discount rates per period. In contrast, behavioral economists note that people often exhibit a "*bias* towards the present" in that they prefer the present more than what would be expected from a rational choice perspective. To be more precise, instead of discounting consistently over time, they apply very high discount rates per period for outcomes in the near future but falling discount rates the further in the future the outcomes arise. This has been termed "hyperbolic discounting" (Phelps and Pollak 1968). It constitutes myopia in that people who are prone to hyperbolic discounting consume more in the present than what they had previously planned. Thaler (1991) illustrated the phenomenon by noting that most individuals would inconsistently prefer one apple today to two apples tomorrow but

two apples in 51 days to one apple in 50 days. Crimes associated with drug addiction provide illustrative examples for hyperbolic discounting. If punishment does not follow swiftly, it will not work as a deterrent because addicts attribute a very high value to the fulfillment of their immediate cravings but a very low value to future costs. In the case of strong addictions, juvenile delinquency, and extreme social deprivation, hyperbolic discounting may be so strong that deterrence based on punishment, which is inevitably separated from the offense by a certain time lag, will not work at all. The stronger the presumptive offender's hyperbolic discounting, the swifter the punishment must follow to work as a deterrent.

Emotions: Many emotions arise from people's social interactions and the apparent human need to reciprocate good as well as bad experiences. Emotional behavior can be related to myopia and hyperbolic discounting in that people succumb to immediate behavioral urges without being able to judiciously consider the future consequences and the legality of their actions. Examples are crimes of passion – directed against individuals by whom the offender believes to have been wronged – such as violent acts of rage or revenge committed at the height of an emotional crisis. The analytical potential of economics of crime, which is inherently based on the assumption of *purposive* action, is limited in the case of spontaneous acts committed in the heat of the moment. But not all emotional crimes are undeliberated acts caused by failing self-control. Instead, some crimes of passion are planned well in advance. The same holds for hate crimes that are directed against people for no other reason than their belonging to certain groups (homosexuals, Jews, people of color, refugees, etc.). People in democratic countries with a high level of economic prosperity are not immune to hate crimes. For example, refugees are often met with anxieties and fears that are amplified by social media echo chambers and exploited by populist politicians. In a hostile public atmosphere, members of deviant political subcultures may feel invited to even physically attack refugees and volunteers without prior reason.

Limited attention: Some noncompliance instances are due to the fact that people make mistakes because they are forgetful and because they can only pay attention to a limited number of issues at a time. If someone focuses strongly on one matter, other issues may receive too little or no attention despite being relevant (Simons and Chabris 1999). Not paying taxes in time or not complying with the complex legal codes of practice in food production, for example, is often not the result of purposive action but the result of negligence caused by the individual's limited attentiveness. This does not rule out an economic approach. Negligence can often be economically deterred because the rule addressees can and will use additional resources and manpower to avoid negligent mistakes if they are more tightly controlled and sanctioned or made liable for damages.

Cross-References

- ▶ Behavioral Law and Economics
- ▶ Bounded Rationality
- ▶ Crime and Punishment (Becker 1968)
- ▶ Crime: Economics of, the Standard Approach
- ▶ Criminal Sanctions and Deterrence
- ▶ Economic Analysis of Law
- ▶ Endowment Effect
- ▶ Experimental Law and Economics
- ▶ Externalities
- ▶ Harmonization: Legal Enforcement
- ▶ Hate Groups and Hate Crime
- ▶ Intrinsic and Extrinsic Motivation
- ▶ Law and Economics, History of
- ▶ Nudge
- ▶ Posner, Richard
- ▶ Prosocial Behaviors
- ▶ Protective Factors
- ▶ Public Enforcement
- ▶ Rationality
- ▶ Retributivism

References

Ayres I, Braithwaite J (1992) Responsive regulation: transcending the deregulation debate. Oxford University Press, New York

Beccaria C (1764) Dei delitti e delle pene. English edition: Bellamy R (ed) (1995) On crimes and punishments and

other writings (trans: Davies R et al.). Cambridge University Press, Cambridge
Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217
Bentham J (1789) An introduction to the principles of morals and legislation. Clarendon Press, Oxford
Braithwaite J (2002) Rewards and regulation. J Law Soc 29(1):12–26
Braithwaite V (2009) Defiance in taxation and governance: resisting and dismissing authority in a democracy. Edward Elgar, Northampton
Dhami S (2016) The foundations of behavioral economic analysis. Oxford University Press, Oxford
Ehrlich I (1975) The deterrent effect of capital punishment: a question of life and death. Am Econ Rev 65(3): 397–417
Frey BS (1997) Not just for the money. An economic theory of personal motivation. Edward Elgar Publishing, Cheltenham
Gigerenzer G (2010) Moral satisficing: rethinking moral behavior as bounded rationality. Top Cogn Sci 2(3):528–554
Gigerenzer G, Todd PM (1999) Simple heuristics that make US smart. Oxford University Press, New York
Gneezy U, Rustichini A (2000) A fine is a price. J Leg Stud 29(1):1–17
Harcourt BE (2014) Beccaria's on crimes and punishments: a mirror on the history of the foundations of modern criminal law. In: Dubber MD (ed) Foundational texts in modern criminal law. Oxford University Press, Oxford, pp 39–59
Harel A (2014) Criminal law as an efficiency-enhancing device: the contribution of Gary Becker. In: Dubber MD (ed) Foundational texts in modern criminal law. Oxford University Press, Oxford, pp 297–316
Hirschauer N, Scheerer S (2014) Protective factors. In: Marciano G, Ramello GB (ed) Encyclopedia of law and economics. Springer, New York
Hobbes T (1651) Leviathan: or the matter, forme, & power of a common-wealth ecclesiasticall and civilll (ed: Shapiro 2010). Yale University Press, New Haven
Kahneman D (2011) Thinking, fast and slow. Farrar, Strauss and Giroux, New York
Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47(2):263–291
Katona G (1953) Rational behavior and economic behavior. Psychol Rev 60(5):307–318
Kolm SC (1973) A note on optimum tax evasion. J Public Econ 2:265–270
Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. Calif Law Rev 88(4): 1051–1144
Kovandzic TV, Vieraitis LM, Boots DP (2009) Does the death penalty save lives? New evidence from state panel data, 1977 to 2006. Criminol Public Policy 8(4):803–843
Markowitz H (1952) The utility of wealth. J Polit Econ 60(2):151–158
Miron AM, Brehm JW (2006) Reactance theory – 40 years later. Z Sozialpsychol 37:9–18
Nielsen VL, Parker C (2012) Mixed motives: economic, social, and normative motivations in business compliance. Law Policy 34(4):428–462
Ostrom E (2005) Understanding institutional diversity. Princeton University Press, Princeton
Phelps ES, Pollak RA (1968) On second-best national saving and game-equilibrium growth. Rev Econ Stud 35:185–199
Picciotto S (2002) Introduction: Reconceptualizing regulation in the era of globalization. J Law Soc 29(1): 1–11
Popitz H (1968) Über die Präventivwirkung des Nichtwissens. Dunkelziffer, Norm und Strafe. Mohr, Tübingen
Posner RA (2014) Economic analysis of law. Wolters Kluwer Law & Business, New York
Radbruch G (2011) Rechtsphilosophie. In: Dreier R, Paulson SL (eds) Studienausgabe, 2nd edn. C.F. Müller, Heidelberg
Rubinstein A (1991) Comments on the interpretation of game theory. Econometrica 59(4):909–924
Sandel MJ (2012) What money can't buy – the moral limits of markets. Penguin, London
Simon HA (1955) A behavioral model of rational choice. Q J Econ 69(1):99–118
Simon HA (1957) Models of man: social and rational. Wiley, New York
Simons DJ, Chabris CF (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception 28(9):1059–1074
Thaler RH (1991) The psychology of choice and the assumptions of economics. In: Thaler RH (ed) Quasi-rational economics. Russell Sage Foundation, New York, pp 137–166
Thaler RH (1999) Mental accounting matters. J Behav Decis Mak 12(3):183–206
Thaler RH (2016) Behavioral economics: past, present, and future. Am Econ Rev 106(7):1577–1600
Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven
The Behavioural Insights Team (2011) Behavioural insights team annual update 2010–11. Cabinet Office, London
Tversky A, Kahneman D (1981) The framing of decisions and the psychology of choice. Science 211(4481): 453–458
Von Neumann J, Morgenstern O (1944) Theory of games and economic behavior. Princeton University Press, Princeton
Weinstein ND, Klein WM (1996) Unrealistic optimism: present and future. J Soc Clin Psychol 15(1):1–8
Zimmerman PR (2004) State executions, deterrence, and the incidence of murder. J Appl Econ 7(1): 163–193

Crime: Economics of, the Standard Approach

Paolo Buonanno[1] and Juan F. Vargas[2]
[1]Department of Management, Economics and Quantitative Methods, University of Bergamo, Bergamo, Italy
[2]Department of Economics, Universidad del Rosario, Bogota, Colombia

Definition

Economics of crime aims at studying, theoretically and empirically, which are the determinants of criminal behavior and how it is affected by incentives and punishment. In 1968, Becker presents a paper that radically changes the way of thinking about criminal behavior. Since the beginning of 1980s, Becker's paper opens the door to a new field of empirical research whose main purpose is to verify and study the economic variables that determine criminal choices and behaviors of agents.

Introduction

In 1968 Gary Becker published "Crime and Punishment: An Economic Approach," a paper that radically changed the economic approach to analyzing criminal (as well as all types of illegal) behavior. Criminal choice conduct is not determined by mental illness or bad attitudes. Rather, it is an individual's *choice*, based on a maximization problem in which agents compare the costs and benefits of misbehavior. The costs are given by the probability of being arrested, the likely punishment, and the gain that is forgone by engaging in a crime instead of in legitimate market opportunities. The benefit is the expected return from committing the crime. Thus, in the Beckerian framework criminal decision-making responds to an economic analysis of agents.

The development of this "rational" approach to criminal behavior, in economics and other social sciences, was favored by "redistributive" and "utilitarian" theories of crime, proposed respectively by Cesare Beccaria (Beccaria 1819) and Jeremy Bentham (1789). In "On Crimes and Punishments," Beccaria argued that crime represents a violation of the social contract, and thus punishment is justified only to defend the social contract and to ensure that everyone will abide by it, deterred from engaging in criminal activities. In turn, in "An Introduction to the Principles of Morals and Legislation," Bentham – largely influenced by Beccaria, first introduced the idea that crime is the consequence of a cost-benefit analysis. He writes: "The profit of the crime is the force which urges man to delinquency: the pain of the punishment is the force employed to restrain him from it. If the first of these forces be the greater, the crime will be committed; if the second, the crime will not be committed" (p. 33). The imprint of the two philosophers on Becker is conspicuous.

The first economics studies of crime came, however, over 200 years after the seminal writings of Beccaria and Bentham. In the 1960s, before Becker's "Crime and Punishment," Belton Fleisher explored the relationship between unemployment and youth crime (Fleisher 1963), and the role of income on the decision to engage in criminal activities (Fleisher 1966). Fleisher was the first economist to analyze empirically the relationship between crime and economic and social variables. However, it was Becker (1968)'s general theoretical framework of crime as an individual's rational choice, what constituted the starting point for analyzing criminal behavior – as well as crime control policies – from an economics perspective.

In his 1993 Economics Nobel Prize acceptance lecture, Becker underlies that while "[i]n the 1950s and '60s, intellectual discussions of crime were dominated by the opinion that criminal behavior was caused by mental illness and social oppressions, and that criminals were helpless victims," the economics approach "[i]mplie(s) that some individuals become criminals because of the financial and other rewards from crime compared to legal work, taking account of the

likelihood of apprehension and conviction, and the severity of punishment." (p. 5).

We now briefly review Becker's basic model of how individuals choose between legitimate and illegitimate activities, basing their decision on a cost-benefit analysis. In Becker's own words "[t]he approach (...) follows the economists' usual analysis of choice and assumes that a person commits an offence if the expected utility to him exceeds the utility he could get by using his time and other resources at other activities. Some persons become 'criminals', therefore, not because their basic motivation differs from that of other persons, but because their benefits and costs differ" (p. 176).

Specifically, Becker defines a supply of offences (O), which relates "the number of offences by any person to his probability of conviction (p), his punishment if convicted (f) and a portmanteau variable (u), such as the income available to him in legal and other illegal activities":

$$O_j = O_j\left(p_j, f_j, u_j\right)$$

An agent's choice is made under uncertainty, then the *expected* utility from committing a crime is defined as:

$$EU_j = p_j U_j\left(Y_j - f_j\right) + \left(1 - p_j\right) U_j\left(Y_j\right)$$

where Y_j the income ("monetary plus psychic") of committing a crime, and f_j "the monetary equivalent of the punishment." The expected utility is also determined by the probability of getting away with the crime: $1-p_j$. The supply of crime is thus decreasing p and f.

To Becker, optimal policies to combat illegal behavior are those that minimize the "social loss" that crime entails. He defines the social loss function from offences as:

$$L = D(O) + C(p, O) + bfpO$$

where D is damage from crime, C the cost of apprehension and conviction, and $bfpO$ is the total social loss from punishments. Here, the social policy variables are represented by p (the probability of arrest) and f (the punishment). The parameter b aggregates individual offender costs of punishment into a social cost. Solving the model with respect to the policy instruments p and f, while taking into account the equilibrium supply of offences, Becker obtains important implications on agents' propensity toward risk. Crime reduction can occur through reducing the benefits of crime, or else from raising the probability of being caught or the costs of punishment conditional upon being caught. Also "a rise in the income available in legal activities or an increase in law-abidingness due, say, to 'education' would reduce the incentive to enter illegal activities and thus reduce the number of offenses" (p. 177).

In 1973, Isaac Ehrlich extended Becker's model, allowing individuals to allocate time between legal and illegal market activities, as well as in nonmarket activities. While Becker's agents are either criminals or not at all (given their cost-benefit calculation), Ehrlich models more sophisticated individuals, in the sense that they can spend time in different activities, both legal and illegal.

In spite of the similarities between Becker's basic model and Ehrlich pioneer (Ehrlich 1973), and subsequent (Ehrlich 1975, 1996) refinements, an important contrast is Ehrlich (1981)'s distinction between the "deterrence" effect of criminal sanction (considered by Becker the sole role of policing and incarceration) and the "incapacitation" effect that locking-up criminals has on crime reductions, by impeding felons to rejoin the crime industry once they are released. In Ehrlich's own words, "deterrence essentially aims at modifying the 'price of crime' for all offenders, potential and actual. (...) [I] ncapacitation, in contrast seek(s) to remove a subset of convicted offenders from the market for offences either by relocating them in legitimate labor markets, or by excluding them from the social scene for prescribed periods of time" (p. 311).

Theoretical refinements of this sort can only improve the design of policies for crime

prevention and crime reduction. After all, the primary objective of Becker himself was to "...use economic analysis to develop optimal public and private policies to combat illegal behavior" (p. 207). Today, the "economics of crime" literature pays constant tribute to the father of the field, with a growing and vibrant set of applications of the basic cost-benefit, rational approach to individual's illegal behavior.

Cross-References

- ▶ Becker, Gary S.
- ▶ Crime and Punishment (Becker 1968)
- ▶ Crime, Attitude Towards
- ▶ Crime: Economics of, Different Paradigms
- ▶ Crime, Incentive to
- ▶ Criminal Sanctions and Deterrence
- ▶ Economic Analysis of Law

References

Beccaria C (1819) An Essay on Crime and Punishments, second American edition published by Philip H Nicklin. Translated from the Italian: "Dei delitti e delle pene" (published in 1764 by Marco Coltellini)

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217

Becker GS (1993) Nobel lecture: the economic way at looking at behavior. J Polit Econ 101(3):385–409

Bentham J (1789) An introduction to the principles of morals and legislation, Oxford: Clarendon Press, 1907 (first published in 1789)

Ehrlich I (1973) Participation in illegitimate activities: a theoretical and empirical investigation. J Polit Econ 81(3):521–565

Ehrlich I (1975) On the relation between education and crime. In: Juster FT (ed) Education, income and human behavior. McGraw-Hill, New York, pp 313–337

Ehrlich I (1981) On the Usefulness of Controlling Individuals: An Economic Analysis of Rehabilitation, Incapacitation, and Deterrence. Am Econ Rev 71(3): 307–22

Ehrlich I (1996) Crime, punishment, and the market for offenses. J Econ Perspect 10(1):43–67

Fleisher B (1963) The effect of unemployment on juvenile delinquency. J Polit Econ 71(6):543–555

Fleisher B (1966) The effects of income on delinquency. Am Econ Rev 56(1/2):118–137

Crime: Expressive Crime and the Law

Martin A. Leroch
Politics and Economy, Johannes Gutenberg Universität Mainz, Mainz, Germany

C

Abstract

Expressive crime contrasts with instrumental crime in that delinquents aim to "make a statement", and not to "make a living". This difference in motivation has important consequences for the deterrent effect of policies. Whereas policies aiming at expected material costs of perpetrators have often proven successful in deterring instrumental crime, they may fail to deter delinquents with expressive motivations. In some cases, perpetrators may even feel defied to increase their activity.

Definition

Expressive crime is a form of illegal behavior which is motivated by the desire to communicate personal attitudes to others. It contrasts with instrumental crime, which is motivated by the desire to gain material objects.

Crime (Expressive) and the Law

Crime may be distinguished along several dimensions. One such dimension focuses on the motivation underlying illicit actions. Accordingly, while instrumental crime is motivated by the desire to "make a living," expressive crime is motivated by the will to "make a statement." Examples of instrumental crime include robbery, stealing, and fraud, while suicide terrorism, illegal political protests, or the spraying of graffiti constitute examples of expressive crime. Some incidents of crime share both components. For instance, severely beating up a debtor who owes money to a criminal organization may on the one hand be regarded a signal or statement to others

that they better repay their debt. On the other hand, this message per se is instrumental in acquiring more riches.

Distinguishing crime according to its motivation is important for the design of policies to deter criminals, one of the major foci of economists, as has been argued, for instance, by Cameron (1988), Kirchgässner (2011), or Robinson and Darley (2004). In theory, deterrence can be achieved either by increasing the probability of detection or by reducing the utility levels of detected delinquents. For the case of instrumental crime, this theory appears valid, at least in parts. While there is evidence that policies aiming at increasing detection probabilities may be successful (e.g., in Klick and Tabarrok 2005), most studies call into question the deterrent effect of increases in available punishment. In cases of expressive crime, however, an increase in detection probabilities or punishment may even lead to perverse effects of increases in both number and intensity of illicit behaviors. The reason for these effects is that punishment may promote crime by affecting "the values of certain other variables . . . which in turn have a direct effect on deviant behaviour" (Opp 1989, p. 421). Crucial for this crime-promoting effect are indirect effects of punishment, which change the (social) incentive structures for perpetrators. Opp finds support for the defiance of perpetrators in examples of legal and illegal political protest, two prominent forms of expressive behavior.

What makes expressive crime so susceptible for these indirect effects of punishment is its signaling property. By making their illegal statements, perpetrators want to send signals to others, for instance, their peers or their opponents. For instance, perpetrators may wish to signal their opponents that they are dissatisfied with the current organization of society or with certain decisions of the current government. Likewise, they signal their active peers that they do not "stand alone" with their attitude. Because the costs associated with transmitting the signal affect its "quality" positively, delinquents may choose to send more (and/or stronger) signals when harsher means of deterrence are employed. Deliberately facing armed forces on the street, for instance, may transmit the signal that I am willing to incur physical pain in order to show others that I disagree with certain policies of the current government. The more likely physical pain is, or the more pain I am to expect, the better I can signal my disagreement. The willingness to incur this pain may also increase my standing among peers – assuming that all peers share opposition toward the government and agree that illegal protest is an adequate form of expressing this opposition.

If standard means of deterrence prove counter-effective, the question of how to counter expressive crime evolves. Addressing the signal value of illicit behaviors or the signal as such appears the most promising strategies. The fight against graffiti in New York City's underground provides a vivid example for such a strategy. After being considered a serious problem in the 1970s, the city managed to almost entirely free its subways from graffiti by 1989. Apparently, the crucial change in strategy was to prohibit sprayed trains from running. Any train leaving the depot was cleaned from graffiti entirely before being brought to service. Sprayers thus knew that spraying was no effective way of transmitting their signals any longer (sprayers typically want to signal disagreement with capitalist ways of organizing society, express artistic desires, or gain attention from others. See Leroch (2014) for a more detailed analysis of graffiti spraying), and they stopped spraying on subways. Cities around the globe copied this strategy successfully in the fight against graffiti.

References

Cameron S (1988) The economics of crime deterrence: a survey of theory and evidence. Kyklos 41: 301–323

Kirchgässner G (2011) Econometric estimates of deterrence of the death penalty: facts or ideology? Kyklos 64(3):448–478

Klick J, Tabarrok A (2005) Using terror alert levels to estimate the effect of police on crime. J Law Econ 48(2):267–280

Leroch M (2014) Punishment as defiance: deterrence and perverse effects in the case of expressive crime. CESifo Econ Stud. https://doi.org/10.1093/cesifo/ift009

Opp K-D (1989) The economics of crime and the sociology of deviant behaviour: a theoretical confrontation of basic propositions. Kyklos 42(3):405–430
Robinson PH, Darley JM (2004) Does criminal law deter? A behavioural science investigation. Oxf J Leg Stud 24(2):173–205

Crime: Organized Crime and the Law

Matthew J. Baker
Hunter College and the Graduate Center, CUNY, New York, NY, USA

Abstract

This entry reviews the literature on the economics of organized crime. While the economics of organized crime is a small subfield of economics, it can offer insights in unexpected areas of economics. The field began some 40 years ago with the study of organized crime and its participation in illicit activities such as prostitution and gambling. Over time, the economic analysis of organized crime has expanded to address broader questions of governance in the absence of formal institutions.

Introduction and Background

The economic literature on organized crime constitutes a rather small portion of the field of law and economics. That having been said, it offers some interesting and unexpected insights into questions of interest to economists. Since its beginnings 40 or so years ago, economic research on organized crime has expanded from a somewhat narrow focus on illicit markets into a broad study of patterns in organizational formation and collective decision-making. The result is that economists' study of organized crime has contributed to the understanding of things of fundamental importance, such as the economics of governance and the formation and organization of property rights.

The phenomenon of organized crime first attracted the interest of economists in the late 1960s and early 1970s. The interest can be attributed to enthusiasm for Becker's (1968) development of an economic theory of criminal behavior and decision-making and also to policy interests deriving from the 1967 report of the President's Racketeer Influenced and Corrupt Organizations (RICO) task force and ensuing adoption of the Organized Crime Control Act of 1970. Initial work focused mainly on description of industries often associated with organized crime, such as loan-sharking, gambling, prostitution, and narcotics, to name a few (see Kaplan and Kessler (1976) for several illuminating examples). Schelling (1967, 1971) provided the first theoretical apparatus for thinking about organized crime. From his initial theoretical steps, the definition and study of organized crime has evolved toward the view that the criminal organization is in fact a form of nascent government, which often arises to perform the usual government functions of enforcement and policing in sectors of the economy where the government is either unable or unwilling to do so.

Defining and Theorizing About Organized Crime

Activity-Based Analysis

Defining organized crime is a difficult task, and, as noted by Fiorentini and Peltzman (1995b, Chapter one) in their review of the literature, research on organized crime has to some degree been driven for a search for an adequate definition of the subject (incidentally, the publication of Fiorentini and Peltzman (1995a) was something of a watershed moment in the economics of organized crime. It is required reading for anyone interested in the economics of organized crime, as is Fiorentini (2000). In fact, this entry has been written with the twin objective of covering necessary ground yet complementing these earlier works in mind). Accordingly, their characterization serves as a useful starting point for characterizing the economic literature on organized crime (defining organized crime is also a difficult multidisciplinary puzzle. See Finckenauer (2005)). As alluded to in the "Introduction" and as Fiorentini

and Peltzman (1995b, Chapter One) emphasize, the first attempts at the definition of organized crime were activity based. A criminal organization was deemed to be one that operates either in full or in part in illegal or illicit markets such as gambling, loan-sharking, and prostitution. Analysis focused on description of the growth and evolution of organized crime in particular industries, operational details, and estimation of the overall size of the industry. One might include in this category even earlier descriptions of the growth and expansion of organized crime, such as analyses of the rise of the American Mafia during Prohibition or analyses of the growth and organization of the Sicilian Mafia (see, e.g., Cressey (1969), who provides a thorough description of the history and nature of organized crime in the United States). One illustrative example of a lucid industry description is given by Kaplan and Matteis (1976) who discuss the day-to-day workings of a loan-sharking operation. Much of this early literature emphasizes that the criminal organization tends to provide goods and services which are illegal, but for which there is nonetheless consumer demand, such as narcotics or prostitution (one might maintain that the development and growth of the Mafia has less to do with serving illegal markets and more to do with filling a vacuum of power. In this way, early literature on the Mafia might have more in common with the theories based upon the development of property rights under anarchic circumstances, which are discussed in section "Governance and Crime"). Kaplan and Matteis (1976), to continue the example, note that loan sharks are specialists in offering credit to an unserved segment of the market: high-risk, short-term borrowers. Of course, description of industry is essential to its study, and this tradition is, thankfully, still very much alive in the economic literature (see, e.g., Bouchard and Wilkins (2010)).

Extortion and Crime

Schelling (1967, 1971) represents the initial break with the activity-based approach to the study of organized crime and also is the first to present a cohesive theoretical structure for thinking about the economics of criminal organizations. His point of departure is that a criminal organization is in essence an extortative body. As such, its primary goal is to maintain monopoly control of illicit activities (and maybe even legitimate ones) for the purposes of rent extraction – protection money. There is no real reason, according to Schelling, that prostitution or gambling or other illicit activities have to be monopolized, but they are obvious targets for extortion because they need some degree of visibility to operate and benefit from regular customer bases, but could not operate as effectively without some shield from the law. Moreover, industries like prostitution and gambling are easily monitored for purposes of rent extraction, as things like income and provider ability are easily monitored.

Schelling's view is attractive in that it has some ability to explain other aspects of organized crime, such as why organized criminal organizations so often engage in turf wars. Buchanan (1973) parlayed Schelling's position into a supplementary theoretical argument offering some good news for society: since illicit activities are by and large those that society deems unacceptable (social "bads"), monopolization of these activities by criminal organizations actually produces a social benefit, as these goods will tend to be underprovided. Backhaus (1979), however, takes exception with this position, noting that, among other things, there are likely to be scale economies in organizing illicit activities. Thus, a monopolist might provide more illicit goods and services to the market.

The first big challenge to the theory evinced by Schelling came from Reuter (1983), who analyzed a broad array of empirical and historical evidence on organized crime and found that Schelling's approach did not always match the facts. In particular, Reuter (1983) emphasized that much organized criminal activity seems to be more competitive than monopolistic (or perhaps at least monopolistically competitive), in that many small, independent entities often provide illicit goods or services in the same market area. Some years later, technical aspects of this observation were taken up by Fiorentini (1995), who developed a former mathematical model of oligopolistic competition in illegal markets. He finds that it

is difficult to draw exact conclusions about how resources invested in violence, corruption, or provision of the goods depend upon the degree of competition and that the answer depends in a critical way on what sort of response criminal organizations expect from the government. This result in fact foreshadows results emanating from recent research on organized crime, governance structures, and imperfect enforcement of property rights. A model in a similar vein is presented by Mansour et al. (2006). They ask how it could be the case that, while deterrence expenditures increased throughout the 1980s and 1990s in the United States, criminal output (e.g., output of drugs) increased and prices fell. The essential idea derives from allowing market structure to respond to the pressures of deterrence; if deterrence creates more competition, output may well expand.

An alternative tradition that also departs from the theory proposed by Schelling, owing originally to Anderson (1979) (see also Anderson (1995)), is a transaction cost approach to studying organized crime. Anderson, in her description of the development and organization of the Sicilian Mafia, argues that transaction costs are important drivers of organized criminal activity. Dick (1995) provides a more expansive transaction cost analysis, which emphasizes what even casual observation suggests are universally important practical problems presented by illicit exchange, such as commitment, credibility, and maintenance of contracting arrangements and long-term relationships. Polo (1995) presents a formal theoretical model capturing credibility and commitment in a long-term setting that bears resemblance to models of contracting and commitment in settings in which agents cannot rely on formal institutions, as in Clay (1995) and Greif (1993). Turvani (1997) also attacks the problem of organized crime from a transaction cost economics perspective.

Governance and Crime

In the early 1990s, an alternative way of thinking about the organized criminal organization as a kind of nascent state emerged. This theoretical approach grew out of the literature on endogenous development and enforcement of property rights in conditions of anarchy, usually associated with the initial efforts of Skaperdas (1992), Hirshleifer (1988), and Grossman and Kim (1995). One essential idea of this work is that security is, at the end of the day, a good like any other, and while the government typically provides it, when it does not, some other actor will invest in the provision of security. Milhaupt and West (2000) provide some data-based empirical support for the idea that illicit activity expands due to inefficiencies in state provision of property rights. Skaperdas and Syropoulos (1995) present a model in which rival gangs allocate resources to both production and security of what is produced. Grossman (1995) models a situation in which a criminal organization arises as a "rival kleptocrat" to the government; that is, the Mafia comes about as a rivalrous organization defending property rights and extracting resources from producers, just as the government imposes taxes. One interesting aspect of Grossman's argument is that the populace might actually benefit from the competition in services provided by the criminal organization. It both expands the amount of protective services available and also limits rent extraction by the licit government.

This view of the criminal organization was subsequently further developed and expanded upon. Skaperdas (2001) provides an expansive view of historical evidence through this lens. At the same time, more sophisticated views of markets and interactions have been applied to the study of organized crime as well as the basic idea that the criminal organization operates in conjunction with the government and is to some degree fueled by the inadequacy of bureaucrats and provision of basic government services. For example, Kugler et al. (2005) present a model in which criminal organizations compete in a "global" market in providing products, but intercede only locally in bribery of bureaucrats. They find increased deterrence could, in some circumstances, lead to higher crime.

Indeed, this branch of literature has evolved toward the position that the governance and security problems which criminal organizations arise to deal with are not all that different from the

problems of social organization faced by human populations throughout history. Skaperdas and Konrad (2012) provide a model in which self-governance is possible and indeed a best-case scenario for the populace. It is, however, less stable and susceptible to more predatory styles of government, and unfortunately, the more competition in the market for protection, the worse things become. Similarly, Baker and Bulte (2010) focus on the Viking epoch beginning around 800 AD and describe how Viking raids – which might be thought of as organized criminal activity – became more and more sophisticated, even as measures to prevent raiding grew more sophisticated.

Internal Structure, Participation, and Deterrence Policy

The basic facts underlying the internal structure of criminal organizations are ingrained in popular myth, yet deeper details and analysis have been a bit harder to come by (of course, much of the work cited in this entry contain institutional detail. See, e.g., Skaperdas (2001) and Anderson (1979)). There is a large empirical and theoretical literature on the decision to participate in crime dating from Becker's (1968) paper, but the decision to participate specifically in organized crime has received considerably less attention.

The notable exception is Levitt and Venkatesh (2000), who analyze in detail the finances and internal structure of a Chicago drug-selling gang. The gang analyzed by Levitt and Venkatesh has a hierarchical structure with a highly skewed wage scale. Those at the bottom of the hierarchy are paid minimal fixed wages for what is dangerous and risky work. Those at the top of the hierarchy do considerably better, motivating Levitt and Venkatesh to suggest promotion and wages are to some degree governed by a rank-order tournament. Still, they find it is difficult to reconcile gang participation with optimizing behavior if one only considers pecuniary rewards from participation. A further finding of Levitt and Venkatesh is that those in the lower echelons of the gang often simultaneously participate in legitimate markets; this suggests that members of the lower rungs might be peeled off by targeted policy interventions.

Garoupa (2000, 2007) explicitly focuses on the internal, vertical structure of the criminal organization using principal-agent theory. Garoupa (2000) follows a more or less classic model of enforcement and deterrence, but allows that offenders must pay a fee to a monopolist agent (i.e., the Mafia) to commit a crime. A key result is that less severe enforcement policy may be optimal than it would be in the absence of a Mafia. Garoupa (2007) works in the same framework, but focuses on a wider assortment of potential punishments and different dimensions of the organization, such as how many agents the organization will seek to hire and how mistake-prone the organization will be in light of its size and, ultimately, the sort of punishments it faces. One result of interest is that severe punishment may reduce organizational scope, but increase organizational effectiveness.

Conclusions

While the economic literature on organized crime remains small, within this literature, a considerable measure of historical detail and methodological diversity has been used in its study. Perhaps the most striking thing about the literature on organized crime is that the very different methodologies employed in its study all recommend a degree of care in thinking about its prevention and deterrence. The recurring theme is that heavier investment in policing and punishing organized criminal activity may have unexpected and unpleasant consequences. The sophistication, competence, and extent of the organization may respond in ways contrary to intuition. In the current domestic and international atmosphere, illicit market activity in the form of international drug and human trafficking remains a major concern, as do the terrorist and other non-state organizations which arise whenever and wherever in the world there is a vacuum of power. In light of this state of affairs, the key insights offered by the literature on organized crime seem to be well worth emphasizing.

References

Anderson AG (1979) The business of organized crime: a Cosa Nostra family. The Hoover Institution, Stanford

Anderson AG (1995) Organized crime, mafia, and governments. In: Fiorentini G, Peltzman S (eds) The economics of organized crime. Cambridge University Press, Cambridge, UK, pp 33–54

Backhaus J (1979) Defending organized crime? A note. J Leg Stud 8:623–631

Baker MJ, Bulte EH (2010) Kings and Vikings: on the dynamics of competitive agglomeration. Econ Gov 11:207–227

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76:169–217

Bouchard M, Wilkins C (eds) (2010) Illegal markets and the economics of organized crime. Routledge, London/New York

Buchanan J (1973) A defense of organized crime. In: Rottenberg S (ed) Economics of crime and punishment. American Enterprise Institute, Washington, DC, pp 119–132

Clay KB (1995) Trade without law: private-order institutions in Mexican California. J Law Econ Org 13:202–231

Cressey DR (1969) Theft of the nation: the structure and operations of organized crime in America. Harper and Row, New York

Dick AR (1995) When does organized crime pay? A transactions cost analysis. Int Rev Law Econ 15:25–45

Finckenauer JO (2005) Problems of definition: what is organized crime? Trends Organ Crime 8:63–83

Fiorentini G (1995) Oligopolistic competition in illegal markets. In: Fiorentini G, Peltzman S (eds) The economics of organized crime. Cambridge University Press, Cambridge, UK, pp 274–288

Fiorentini G (2000) The economics of organized crime. In: Bouckaert B, De Geest G (eds) Encyclopedia of law and economics, volume V: the economics of crime and litigation. Edward Elgar, Cheltenham, pp 434–459

Fiorentini G, Peltzman S (eds) (1995a) The economics of organized crime. Cambridge University Press, Cambridge, UK

Fiorentini G, Peltzman S (1995b) Introduction. In: Fiorentini G, Peltzman S (eds) The economics of organized crime. Cambridge University Press, Cambridge, UK, pp 1–30

Garoupa N (2000) The economics of organized crime and optimal law enforcement. Econ Inq 38:278–288

Garoupa N (2007) Optimal law enforcement and criminal organization. J Econ Behav Organ 63:461–474

Greif A (1993) Contract enforceability and economic institutions in early trade: the Maghribi traders' coalition. Am Econ Rev 83:524–548

Grossman HI (1995) Rival kleptocrats: the mafia versus the state. In: Fiorentini G, Peltzman S (eds) The economics of organized crime. Cambridge University Press, Cambridge UK, pp 143–155

Grossman HI, Kim M (1995) Swords or plowshares? A theory of the security of claims to property. J Polit Econ 103:1275–1288

Hirshleifer J (1988) The analytics of continuing conflict. Synthèse 76:201–233

Kaplan LJ, Kessler D (1976) An economic analysis of crime: selected readings. Charles C. Thomas, Springfield

Kaplan LJ, Matteis S (1976) The economics of loansharking. In: Kaplan LJ, Kessler D (eds) An economic analysis of crime: selected readings. Charles C. Thomas, Springfield, pp 178–192

Kugler M, Verdier T, Zenou Y (2005) Organized crime, corruption, and punishment. J Public Econ 89:1639–1663

Levitt SD, Venkatesh SA (2000) An economic analysis of a drug-selling gang's finances. Q J Econ 115:755.89

Mansour A, Marceau N, Mongrain S (2006) Gangs and crime deterrence. J Law Econ Organ 22:315–339

Milhaupt CJ, West MD (2000) The dark side of private ordering: an institutional and empirical analysis of organized crime. Univ Chic Law Rev 67:41–98

Polo M (1995) Internal cohesion and competition among criminal organizations. In: Fiorentini G, Peltzman S (eds) The economics of organized crime. Cambridge University Press, Cambridge, UK, pp 87–103

Reuter P (1983) Disorganized crime: the economics of the visible hand. MIT Press, Cambridge, MA

Schelling TC (1967) Economics and the criminal enterprise. Public Interest 7:61–78

Schelling TC (1971) What is the business of organized crime? Am Sch 40:643–652

Skaperdas S (1992) Cooperation, conflict, and power in the absence of property rights. Am Econ Rev 82:720–739

Skaperdas S (2001) The political economy of organized crime: providing protection when the state does not. Econ Gov 2:173–202

Skaperdas S, Konrad K (2012) The market for protection and the origin of the state. Econ Theory 50:417–443

Skaperdas S, Syropoulos C (1995) Gangs and primitive states. In: Fiorentini G, Peltzman S (eds) The economics of organized crime. Cambridge University Press, Cambridge, UK, pp 61–81

Task Force Report: Organized Crime (1967) The President's commission on law enforcement and administration of justice. U. S. Government Printing Office, Washington, DC

Turvani M (1997) Illegal markets and the new institutional economics. In: Menard C (ed) Transactions cost economics. Edward Elgar, Cheltenham

Crimes Against Nature

▶ Sex Offenses

Criminal Constitutions

Nicholas A. Snow
Department of Economics, Indiana University, Bloomington, IN, USA

Definition

Criminal constitutions exist in a number of different criminal enterprises, from the Golden Age of Pirates to modern prison and street gangs. The purpose of these constitutions is to better facilitate cooperation among organizational members in order to help better achieve profit maximization.

Criminal Constitutions

It is easy to imagine criminal organizations being composed of lawless, all out for their own chaotic individuals but this is far from the truth. From an economic perspective, especially since Gary Becker's famous 1968 publication, *Crime and Punishment: An Economic Approach,* the starting point for analyzing criminal behavior is to assume that criminals are rational agents. From this perspective, it is not surprising to find that many black markets and other criminal organizations are extremely orderly and rational. This is, of course, not to say that they are desirable or morally good but if we are to truly understand criminal organization it is important to analyze them in truth and not in fiction.

Many criminal organizations are not only rational and orderly but also formally created as such. For example in his 1991 book, *Islands in the Streets,* Jankowski found that 22 out of 37 street gangs had written criminal constitutions. And street gangs are only one of numerous examples of criminal organizations having deliberately written and constructed constitutions that attempt to align the self-interests of its members to that of the group interest, such as eighteenth century pirates, prison gangs, the Mafia, etc. Additionally, other black markets without formal constitutions will often still contain informal and implicit criminal codes that exist to do the same thing, such as the market for smuggled liquor in Detroit in the United States in the 1920s under alcohol prohibition as illustrated by historian Larry Engelmann's 1979 book, *Intemperance, the Lost War Against Liquor.*

The reason for the emergence of such criminal constitutions is simple. Black market firms are still firms and as such act in order to maximize profits. Economists Leeson and Skarbek, in a 2010 article, *Criminal Constitutions,* highlight three main reasons for the emergence of criminal constitutions. The first is that constitutions help to create common knowledge about what the organization expects from its members and the different members can expect from each other. This helps to align expectations among participants within the organization and in the broader market setting. Common knowledge is often created by having explicit rules, whether written or not, of how to behave within the organization and without. This has the advantage of members knowing exactly what they are getting themselves into by joining the organization in terms of duties, rewards, obligations to other members, etc. Rules may also consist of entrance requirements for potential members. By knowing the rules when you come into the organization, current members can have more confidence in the new members.

Second, criminal constitutions help to solve the principle agent problem within criminal organizations. It is often the case that members of the various criminal organizations are not necessarily the residual claimants of the organization, thus making them particularly vulnerable to divergent interests. The rules within the constitutions, and their enforcement, can help to make sure that divergence of interests does not take place. For criminal organizations this may be extremely important. For example, as Leeson explains in his 2009 book, *The Invisible Hook,* a pirate's profitable enterprise consists of plundering on the high seas. This is often hazardous work. In the act of successfully plundering a merchant vessel, pirates might be required to

engage in a dangerous battle. Individual crew members may recognize the high costs of participating in battle to themselves. Thus, a collective action problem emerges because the individual pirate's incentive is to stay out of the fight but the more pirates that think like this, the less likely the ship is going to be taken. Profit maximization requires pirates to take as many ships as possible. Therefore, pirates created rules and rewards to encourage all individuals to engage in battle when necessary, such as providing, what essential boiled down to, workman's compensation for injury and special rewards for particular courage.

Finally, criminal constitutions create information about member misconduct and help formulate rules and enforcement that prohibit such behavior. This final function is self-enforcing because conflict within any organization will necessarily cut into the profits of the illicit firm. If members of the gang are constantly fighting, then they are not engaged in productive efforts, which earn the organization profits. As Skarbek illustrates, in his 2014 book, *The Social Order of the Underworld,* prison gangs even utilize rules of conflict for preventing intra-gang conflicts. Violence is costly for business, so the gangs have an incentive to enforce rules among their own members to avoid unnecessary conflict even with competing gangs.

The main distinction facing illicit organizations from legal organizations is the lack of access to formal governance institutions necessary for markets to properly function and flourish. Given the illegal nature of their activities, criminal organizations are unable to turn to legal and social institutions, which help to enforce private property rights, contracts, and facilitate collective action agreements. It is for this reason that most black markets consist of many, very small, and ephemeral enterprises. Thus, contrary to conventional wisdom, large monopolies are often a rare occurrence within black markets. While the use of violence is often unlikely to have an effect on this structure, it is, however, likely to have an effect on profit opportunities for the different organizations within the market. Constitutional arrangements are then able to help facilitate cooperation within and between organizations.

Ultimately, as profit maximizing organizations, criminal organizations use constitutions in order to help facilitate cooperation among its various members to help ensure higher profits for the organization. This certainly seems strange considering these organizations are made up of individuals who are openly flaunting the governments laws but when understanding that just like a legitimate firm, criminal organizations wish to maximize their profits, it becomes clear why such rules would emerge. Markets need rules to function and because of their illicit nature the government is not an option for providing such rules, thus the organizations will create these rules in order to facilitate the gains from trade.

Cross-References

- ▶ Anarchy
- ▶ Crime and Punishment (Becker 1968)
- ▶ Prohibition

References

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217

Engelmann L (1979) Intemperance: the lost war against liquor. Free Press, New York

Jankowski MS (1991) Islands in the street: gangs and American Urban Society. University of California Press, Berkeley

Leeson PT (2009) The invisible hook: the hidden economics of pirates. Princeton University Press, Princeton

Leeson PT, Skarbek D (2010) Criminal constitutions. Glob Crime 11(3):279–298

Skarbek D (2014) The social order of the underworld: how prison gangs govern the American penal system. Oxford University Press, Oxford

Further Reading

McCarthy DMP (2011) An economic history of organized crime: a national and transnational approach. Routledge, New York

Reuter P (1985) The organization of illegal markets: an economic analysis. U.S. Department of Justice: National Institute of Justice, Washington, DC

Criminal Sanctions and Deterrence

J. J. Prescott
Law School, University of Michigan, Ann Arbor, MI, USA

Abstract

This entry defines criminal sanctions by distinguishing them from civil sanctions, and briefly surveys the major categories of criminal sanctions, both ancient and new. The entry then outlines the primary social justifications for using such sanctions—focusing on deterrence as a distinct purpose of punishment—and describes a basic model of criminal offending to highlight the conditions under which the threat of criminal sanctions can influence offender behavior in predictable ways. Next explored are the implications for deterrence of the different types of criminal sanctions. A brief discussion of the suggestive conclusions emerging from related empirical evidence follows.

Definition

Criminal sanctions are punishments (e.g., imprisonment, fines, infliction of pain, or death) imposed by governments on individuals or corporate entities for the violation of criminal laws or regulations. Deterrence, one of the principal theories used to justify the use of criminal punishment, occurs when the possibility that an individual will be subjected to a criminal sanction were he to commit a crime causes the individual to forgo or engage in less of the behavior in question. Deterrence is the primary justification for punishment in the field of law and economics.

Introduction

To address socially harmful (or potentially harmful) behavior, governments enact laws (or groups of people develop practices) to sanction or punish individuals who choose to engage in it. If the problematic behavior rises to a particular level of harmfulness or has particular features (e.g., physical violence), the behavior may be classified as criminal, and upon conviction, a violator will, by definition, receive some sort of criminal sanction. The distinction between a criminal sanction and a civil sanction is largely one of degree and possibly just semantics, although many scholars have attempted to carefully delineate the two spheres. There is a credible argument that a crime is simply a harmful act that the government decides to label a crime (perhaps because it is seriously harmful but perhaps not), and a criminal sanction is simply a sanction the government applies to someone who has performed such an act.

This entry will begin by briefly exploring what may make a sanction "criminal" in nature and then by summarizing the key categories of criminal sanctions. Some types of criminal penalties are largely historical relics; others are innovative and rely on modern technology. Nevertheless, sanctions of more recent vintage, such as constant, real-time GPS monitoring, share many features with more traditional varieties. This entry will primarily focus, however, on deterrence—specifically, its definition and how the imposition of criminal sanctions can achieve this fundamental purpose of punishment. Basic features of the economic model of crime will be described (see Becker 1968), and a variety of topics relevant to criminal deterrence will be covered, including the model's theoretical predictions for offender behavior, the implications for deterrence of employing different categories of modern criminal sanctions (including their unintended consequences and the potential for sanction complementarity when they are used simultaneously), and the consensus conclusions of the empirical literature.

Criminal Versus Civil Sanctions

What makes a sanction a "criminal" sanction? The cleanest answer to this question is the tautological response that "a criminal sanction is a sanction that seeks to punish someone for having committed a

crime," thus begging the question: "What makes a particular act a crime?" One important attribute of a crime is that the act or its consequences be socially harmful. But the category of "harmful" behaviors is ultimately socially constructed through some aggregation of individual preferences and beliefs and is therefore determined simply by whether a government decides that the act in question is indeed harmful enough to be worthy of public censure. Some harmful acts are not subject to government sanctions for practical or philosophical reasons. When a government decides to discipline someone who has engaged in a harmful act, it selects between or some combination of "civil" measures (traditionally, fines) and criminal sanctions.

It is impossible to identify conditions that, without exception, distinguish a crime from another form of civil wrong, like a tort. But there are tendencies, or characteristics, that make a certain harmful act more likely to be labeled a crime. One important criterion is the mental state of the wrongdoer. Crimes are usually (but by no means always) done knowingly or intentionally or at least involve a blameworthy awareness of risk; otherwise innocent acts that accidentally result in harm, however, are rarely considered crimes, and sometimes even intentional acts that cause harm are merely intentional torts (Miceli 2009, p. 269).

Other suggested criteria relate to the difficulties that harmed individuals (or their friends or family) might face (as a group) in privately enforcing particular prohibitions. For example, when particular categories of wrongdoers might be difficult to identify (e.g., thieves), using specialists may make sense, and the average victim is certainly no specialist (Polinsky and Shavell 2007, p. 406). Likewise, if the nature of the harm is widespread, enforcement may be a public good. No single private individual may have the incentive to pursue the wrongdoer. As Miceli (2009, p. 270) points out, extreme examples of this category are "victimless" crimes (e.g., possessing child pornography or drugs) in which the harms are often thought to be the knock-on, indirect consequences of the activity.

Some argue that the civil/criminal wrong divide maps to the liability/property rule distinction in the law (Miceli 2009, p. 273). Under this rubric (see Calabresi and Melamed 1972), civil wrongs are those for which there is, ultimately, a price, usually set by the court ex post in the form of damages awarded in a civil trial. By contrast, the enactment of a criminal prohibition can be viewed as the granting of an inalienable right to potential victims by the government. There is no dollar figure that an offender can pay to rectify the violation of this right. Rather, the offender must be punished, and the punishment has no necessary relationship to (and is usually much greater than) any damages sustained by the victim.

Even better, however, would be the prevention of the harm ex ante, since it is in some sense irreversible. This is consistent with criminal law often being publicly enforced by specialists (Shavell 2004, p. 575), as ex ante prevention by private individuals seems much less likely to succeed relative to winning an ex post lawsuit.

Types of Criminal Sanctions

Typical criminal sanctions in modern countries include fines, incarceration, and supervision (including probation and parole). The death penalty (or capital punishment) is a historically important criminal sanction, but it is employed rarely in practice in Western countries these days and is treated elsewhere in this collection.

Fines. When a fine is imposed, an offender is legally ordered to pay to the state or perhaps to a third party the levied amount, perhaps in lump sum or perhaps in installments under a payment plan. Fines are limited in their effectiveness as a punishment by the assets or potential assets of the offender (and by extension, the offender's ability to work or otherwise access resources) and by the capacity of the government to collect the fine, through garnishment of wages, for example.

Incarceration. Imprisonment or incarceration more generally is perhaps the best-known and most common form of criminal sanction in the modern world, at least with respect to serious crimes. The "quantity" of imprisonment varies by the conditions of the imprisonment (e.g., amount

of space, cleanliness, quantity or quality of food, available or required activities, access to other prisoners, access to loved ones, and so on) and by the time the offender is sentenced to remain imprisoned.

Intermediate Sanctions. Intermediate or alternative sanctions are typically viewed as less serious than incarceration: the offender is released from (parole) or never sentenced to (probation) incarceration but remains free only by compliance with specified conditions, which can be affirmative obligations or negative restrictions.

Traditional forms of intermediate sanctions are increasingly common, at least in some jurisdictions—e.g., home detention, mandated community service, and mandatory drug treatment. Other forms of punishments are considerably less common. For example, rare at least in most Western countries is the imposition of the death penalty, the killing of an individual by the government as punishment for the commission of a crime. Corporal punishment, the deliberate infliction of pain as punishment for a criminal offense (e.g., caning or whipping), is still practiced, although today only in a small minority of countries and with significant limitations.

Governments, including many in the USA, also employ shaming as a criminal sanction—i.e., subjecting an offender to public humiliation as a form of punishment. Historically, these punishments were often accompanied by an element of physical discomfort or of physical or verbal abuse (e.g., stocks or pillory). Modern shaming punishments retain the humiliation element and perhaps allow verbal abuse by the public, but do not typically impose conditions that might result in extraordinary physical discomfort or pain. Finally, banishment, as a form of punishment, is the mirror image of incarceration. Rather than an offender being required to be in a particular place, he is required to leave the community altogether in order to isolate that individual from friends and loved ones and to eliminate any future threat to the community.

Many of the most innovative criminal sanctions in recent years have emerged in response to public concern over sex offenses, especially in the USA. In the 1990s, sex offenders were believed to be highly likely to return to sex crime upon release from incarceration, and many such offenses are serious crimes; as a result, governments developed new criminal sanctions that manage to combine elements of incapacitation, shaming, and banishment.

Sex offender registration, for example, is an extremely weak form of incapacitation. When required to register, offenders must provide identifying information (e.g., name, date of birth, address, criminal history) to law enforcement to facilitate the latter's monitoring and, if necessary, apprehension of them. Consequently, committing a new crime becomes more difficult for an offender who in some sense weighs the costs and benefits of his options. Mandatory real-time GPS monitoring of an offender's location is a more technologically advanced version of registration. If monitoring equipment is visible to others when it is worn, shaming also occurs, much like the proverbial wearing of a scarlet letter.

Community notification is a stronger form of incapacitation, coupled with a different form of shaming. When an offender is made subject to notification, the government releases identifying registry information to the public, which increases the overall level of monitoring, prompts potential victims to take precautions (in theory creating a bubble around the offender), and subjects the offender to humiliation, to great difficulty in finding employment and housing, and potentially to more active forms of abuse at the hands of members of the public. Yet while GPS monitoring equipment may result in humiliation in every one-on-one interaction with another person if the equipment is visible, an individual living under notification, even when the notification is an active form (e.g., neighbors of an offender receive a card in the mail containing the offender's registration information), is not physically marked as a sex offender and so may have anonymous interactions without humiliation.

Sex offender residency restrictions are a targeted form of banishment. Offenders living under residency restrictions are not necessarily forbidden from a particular place under all conditions. However, offenders cannot lawfully live within a certain distance of places where potential victims are particularly likely to frequent. Other types of

restrictions—travel restrictions, employment restrictions, and so on—function in the same way, by banishing offenders either from places where or from roles in which they are assumed to pose the most threat.

Under many criminal justice systems, offenders are sentenced to combinations or bundles of different kinds of criminal sanctions. For example, sentences that include both fines and incarceration are common, as are sentences that begin with incarceration but are followed by parole or some other form of supervision. These combination sanctions may be attempts to pursue multiple purposes of punishment simultaneously (e.g., a fine can punish, but it cannot incapacitate, at least not to the extent that a prison cell can) or they may be an attempt to punish more efficiently. For example, as we will see below, fines can be more efficient as a form of punishment because they are a transfer, but most offenders are liquidity constrained and so fines have only limited utility in most circumstances. Finally, there may be economies of scope in using a cluster of criminal sanctions as a consequence of behavioral tendencies, including, e.g., hyperbolic discounting. Increasing a sentence from 5 to 6 years may be less effective at altering behavior than adding a comparable fine or other restrictions to a 5-year sentence.

Finally, while not formally sanctions, arrest and criminal process by themselves will result in many financial, psychic, and opportunity costs (see Eide 2000, pp. 351–352 for a discussion of these—e.g., lost employment and legal fees), and these costs have the potential to affect an individual's decision to engage in crime (e.g., Bierschbach and Stein 2005).

Purposes of Criminal Sanctions

The application of criminal sanctions to an offender is usually justified in one of two ways. Retributive theories are premised on the idea that, by committing the crime, the offender has become morally blameworthy and is deserving of punishment. A criminal sanction does justice (for society, for the offender, for the victim) by punishing the offender, with the degree of punishment having a direct relationship to the seriousness of the offender's moral culpability (which in turn has some relationship to the seriousness of the harm), at least according to some retributive views. Utilitarian theories, by contrast, are forward looking, focused on future consequences: a criminal sanction is appropriate if the benefits that follow from imposing the sanction outweigh the costs and suffering of its imposition. The punishment that generates the most net benefit is morally preferable. Mixed theories of punishment draw from both sets of ideas, usually with the hope of accounting for our intuitions about whether and how much punishment is appropriate. For instance, certain limited forms of retributivism assume the consequences of criminal sanctions matter at the margin but that moral desert establishes constraints on the levels of punishment that are minimally required and maximally permitted.

Economists are primarily interested in utilitarian or consequentialist theories of punishment, and in setting punishments, they seek to maximize net social welfare by reducing the total harm that results from crime—including the administrative costs and consequences to offenders of imposing criminal sanctions, in addition to the harms suffered by victims of crime and society. In practice, criminal sanctions can reduce future harm in one of three ways, although these categories are not always precisely demarcated: rehabilitation, incapacitation, and most important from the perspective of economists, deterrence.

Rehabilitation involves using criminal sanctions to change the preferences of offenders or to improve an offender's range of legal choices, thereby making the commission of crime less attractive. Incapacitation entails increasing the costs to the offender of making illicit choices or simply limiting the number of illicit choices available (Freeman 1999, pp. 3540–3541). In the latter case, incapacitation can succeed even in the face of an entirely irrational offender. For a discussion of the economics of these purposes of punishment as well as brief notes on the economics of retribution, see Shavell (2004, pp. 531–539).

Deterrence assumes at least some offenders are at least somewhat rational, which means that

they must be capable of responding to incentives and to their environments generally, a proposition that seems obviously true (Shavell 2004, p. 504). When deterrence can work, it is particularly attractive because sanctions need only be employed when an individual breaks the law. In the limit, if deterrence is perfect and there are no law enforcement mistakes, then society can achieve the first best (at least if any benefits to an offender of committing a crime are not included in the calculus): no harm comes to victims and no harm comes to potential offenders. What is more, the nature of the criminal sanction is unimportant, so long as deterrence is complete. Outside of this extreme, however, the relative advantages of different criminal sanctions matter a great deal in determining the socially optimal punishment regime.

Economic Model of Crime and Deterrence

There are at least a few different ways to conceive of developing an economic model of crime. The most straightforward starts with a set of crimes fixed by assumption (e.g., by taking the set of criminal prohibitions enacted by a government as given and simply assuming it includes the harmful acts society ought to criminalize). Analysis then begins with the question: "What steps should the government take to minimize the aggregate net harm of these acts?" To simplify matters, the discussion below proceeds largely using this framing of the problem.

As suggested above, however, a more general approach to the economic model of crime begins first by endogenously identifying specific harmful acts (either because they are harmful in themselves or because they risk harm) that are socially undesirable (Shavell 2004, p. 471). According to the standard economic account, a socially undesirable act is one in which the expected social benefits that result from the act (which may or may not include gains that accrue to the offender) are outweighed by its associated harms. Theoretically, categorizing acts in this way requires comparing the social costs and benefits of the act under all potential enforcement regimes (including no enforcement), making the problem anything but trivial. For certain harmful acts, even the most efficient method of enforcement will prove too costly (Shavell 2004, p. 486); in these cases, minimizing harm is best accomplished through decriminalization, leaving it either to private parties to enforce through civil enforcement mechanisms or to social norms (Polinsky and Shavell 2007, pp. 446–447).

Starting with the categories of acts it considers sufficiently undesirable to criminalize (or with all acts if the set will be endogenously determined), society seeks to minimize the net social harm (or maximize the net benefit) from these, potentially using criminal sanctions in combination with other policies.

The economic model of crime focuses on the individual offender's decision to commit one of these acts, asking whether, given the offender's preferences and environment, his individual benefits of proceeding outweigh his individual costs of doing so (Becker 1968). Most economic models of crime assume that at least some share of potential offenders are capable of responding in predictable ways to incentives created by criminal sanctions (Eide 2000, pp. 352–355). If individuals are incapable of rational behavior in this sense, then enforcement of any kind must be premised either on rehabilitation or incapacitation or both. By assumption, criminal prohibitions are enforced through the probabilistic application of criminal sanctions. The optimization problem is how best to select and structure the application of these sanctions, given their costs and the likely response of relevant actors. Again, this approach may mean allowing certain harmful acts to go unpunished.

Modeling Criminal Behavior

Suppose an individual considers committing a criminal act instead of engaging in some legal (perhaps income-generating) activity. Assume the relative benefits to that individual for carrying out the act *absent* a public enforcement regime (net of the forgone benefits from the competing legitimate activity) amount to b_o. Note that there may also be some social benefit to this act, either b_o (i.e., if society "counts" the gains to the offender and there are no other benefits) or some other lesser or

greater amount, which we may call b_s. By assumption, b_s will not affect the decision of the offender to engage in criminal activity, but b_s may affect whether society ought to label the act a crime as well as how best to enforce any prohibition. If b_o is positive, then by definition the individual will engage in the activity, absent some threat of a criminal sanction.

Now assume a public enforcement regime is in place, and if the offender is caught engaging in the now-prohibited activity, he will be apprehended and criminally sanctioned (e.g., fined or imprisoned for some length of time). This sanction is costly, both to the individual (s_o) and possibly also to society (s_s) (again, s_s may be larger or smaller than s_o).

Public enforcement of criminal prohibitions is infused with uncertainty. Not all criminal activity is detected; if it is detected, not all of it results in an arrest; if the activity results in an arrest, not every arrest results in a conviction that leads to a sanction (Nagin 2013, pp. 207–213). Furthermore, there is the possibility of plea bargaining—i.e., "settlements dilute deterrence," as noted by Polinsky and Shavell (2007, p. 436). Even the formal criminal sanctions announced after conviction are sometimes overturned on appeal or adjusted down the road (e.g., early release). Often, there is also a considerable time lag between the commission of a crime and the application of any sanction to the offender.

At the same time, in some contexts, arrests and criminal process function in effect as additional punishment (conditional only on arrest) because dealing with the police and with criminal allegations is also costly, regardless of whether the offender is ultimately convicted of the crime. At an even more basic level, anxiety over possible detection increases the effective severity of any potential punishment. Criminal sanctions also result in other lost opportunities and collateral consequences not captured by the formal sanction itself (e.g., loss of employment).

Finally, uncertainty also cuts the other way: individuals innocent of any crime may be wrongfully convicted of a crime (Polinksy and Shavell 2007, pp. 427–429), which means even deciding to forgo committing a harmful act does not completely insulate someone from criminal sanctions, although we assume that the probability of a criminal sanction is much higher if the person engages in the prohibited activity.

Let p represent the net increase in the probability that an individual is punished if he chooses to commit the criminal act in question, taking into account all of the considerations outlined above. In effect, p implicitly adjusts for the fact that a criminal sanction ultimately imposed (i.e., experienced) is very different from the one laid out in the law. With certainty, p will not only be less than one (unless detection is extremely high, as it may be for some crimes) but will also be greater than zero, at least in any society in which law enforcement does not arbitrarily mete out punishment. Importantly, p is likely to be specific to an individual; some individuals are better able to avoid detection and conviction, and others are presumably more likely to be wrongfully prosecuted or subjected to more onerous criminal process.

The phrase "deterrence is a perceptual phenomenon" (e.g., Nagin 2013, p. 215) refers to the ideas that (1) the offender must perceive the threat of being criminally sanctioned in order to reasonably expect the offender's behavior to change and that (2) an offender will only be deterred to the extent of the *perceived* level of enforcement or severity of the criminal sanction. Legal systems typically assume that individuals are aware of the law as it appears on the books, and potential offenders typically have some sense of the potential consequences they face for engaging in criminal behavior, but any such knowledge is imperfect (Shavell 2004, p. 481). Allow $f(p \times s)$ to represent a function that captures how potential offenders perceive threatened criminal sanctions and public enforcement more generally. Again, this perception is specific to an individual (Eide 2000, p. 352), as some individuals will be more experienced, better educated, and so on.

These assumptions and this notation allow us to characterize the decision of the potential offender as willing to commit an offense (rather than engage in some other lawful activity) if the net expected utility (relative to alternatives) is positive: $u_o(f_o(p_o \times s_o), b_o) > 0$. Put in the simplest of terms, the potential offender weighs the

net benefits of harmful conduct (absent any threat of enforcement) and compares this to the perceived net individual costs of becoming subject to criminal enforcement. Note that this margin is, by assumption, net of competing legal opportunities that are otherwise available to the individual. Potential offenders who expect to receive relatively large net benefits from the illicit activity, those who underestimate public enforcement efforts, and those who do not expect to suffer as much as others from any sanction, for example, are more likely, all else equal, to choose to offend.

Assuming some level of offender rationality and access to information, the model tells us that society can influence an offender's calculation (and therefore behavior) by changing (1) the level or type of criminal sanction (s_o), (2) how it will be enforced and how likely it is to be enforced (p_o), (3) how the sanction and enforcement are perceived ($f(\cdot)$), and (4) the net benefit (relative to legal pursuits) to the offender of the activity (b_o). Deterrence research traditionally focuses on the first two variables (on the theory that they are more policy relevant), taking into account the potential offender's preferences and other relevant attributes (particularly, risk aversion and wealth or income levels). Reducing the returns to undetected crime also matters in this model, however, and society can always modify potential offender behavior by altering an individual's preferences, although the analysis is agnostic on how a society might accomplish this.

Importantly, a complete theory of optimal public enforcement incorporates this range of strategies and the expected response of offenders to each of them, but also much more that is beyond the scope of this entry's discussion. Identifying how to maximize a social welfare function requires not simply accounting for the effects on offender behavior of criminal sanctions and the degree of enforcement (deterrence). Also vital are the social costs of these activities (Polinsky and Shavell 2007; Miceli 2009), including consequences for victims. Here, the focus is not the optimal enforcement bundle, but deterrence alone—specifically, what models and empirical work can tell us about changing offender behavior by altering criminal sanctions and, to a lesser extent, enforcement levels.

Theoretical Predictions

In a simple deterrence framework, the predictions of how changes in criminal sanctions or enforcement levels affect behavior follow fairly straightforwardly from traditional ideas about consumer choice (or labor supply) and decision making under uncertainty (see, e.g., Freeman 1999). As Eide (2000, p. 345) puts it, deterrence theory is "nothing but a special case of the general theory of rational behavior under uncertainty. Assuming that individual preferences are constant, the model can be used to predict how changes in the probability and severity of sanctions. . . may affect the amount of crime." For the sake of brevity, the discussion below concentrates principally on the potential effects of increasing or decreasing criminal sanctions or the degree of enforcement effort and on the consequences of using different types of criminal sanctions.

To organize the discussion, observe the following: First, as theoretical work makes clear, a potential offender's attitude toward risk (whether risk averse, risk neutral, or risk preferring) has significant implications for criminal behavior in virtually every remotely realistic model of deterrence (e.g., Eide 2000, p. 350). Second, the consequences of different risk attitudes differ depending on the policy lever in question, whether it is the level of enforcement activity, the severity of the criminal sanction, or the type of criminal sanction.

Furthermore, it is important to recognize that theoretical work and empirical evidence on deterrence and offender behavior are most useful in the interior of the policy choice set, not at the extremes. Shavell (2004) has written that "optimal law enforcement is characterized by under deterrence—and perhaps by substantial under deterrence—due to the costliness of enforcement effort and limits on sanctions" (p. 488), implying, if it is not already obvious, that a nontrivial portion of potential offenders will not be deterred in any reasonably practicable regime. At the same time, no one doubts criminal sanctions deter in some basic sense if viewed from the perspective

of lawlessness, whether zero sanctions or no enforcement. As Robinson and Darley have stressed in a number of articles (e.g., 2003, 2004, among others), the relevant policy question is whether criminal sanctions deter effectively *at the margin* or whether long prison sentences deter more effectively *relative to* more moderate but still significant sentences.

Begin by stepping back from the model laid out above and consider perhaps the simplest choice environment available: the decision to allocate a fixed amount of money between legal and illegal activities (see Schmidt and Witt 1984, pp. 151–154). The legal investment provides a certain return; the illegal investment offers a higher return, but the possibility of apprehension (p) and a fine (s) makes it uncertain. The relative returns of these investments vary continuously across investors in such a way as to create two corner solution groups; some proportion makes only legal investments, while others make only illegal investments. Now, consider an increase in p. In all states of the world, the illegal investment offers a weakly lower return, and so under minimal conditions, some proportion of investors near the cutoff move from making illegal investments into making legal investments, regardless of risk preference. Similarly, an increase in the fine (s) also lowers the return on average and weakly in all states, and it does so in a way that generates significantly more risk; so assuming risk aversion or risk neutrality, deterrence also occurs (with fewer illegal investors). These outcomes align roughly with Becker's (1968) conclusions, which suggest that increasing fines rather than enhancing the degree of enforcement will prove more effective at deterring criminal behavior, assuming potential offenders are risk averse.

A more complicated scenario emerges from a somewhat more realistic framework that allows for leisure time as an element of the utility function, even with simplifying assumptions. Imagine a potential offender considering an act enforced by threat of a monetary fine. Assume the crime in question is fraud and the benefit is monetary and that the offender could otherwise spend the time lawfully employed, earning wages with certainty, or enjoying leisure. For ease, assume that u, f, and b and any nonlabor income are fixed and that the initial values of s and p would result in the offender choosing to commit at least some fraud but also realistically engaging in a minimum of some leisure.

What are the effects of increasing either the size of the fine (s) or the degree of enforcement (p)? A portfolio or labor supply analysis provides the answer: because an increase in either s or p would effectively reduce the "return to crime," the prediction is ambiguous. The substitution effect alone will typically point in the direction of less crime (or fewer people engaging in crime), but the income effect may point in the opposite direction, depending on offender attitudes toward risk. More precisely, when the question is which of these two effects dominates, the "tipping point" is determined by the curvature of the utility function (i.e., by the level of risk aversion).

In many deterrence models (Eide 2000, pp. 348–350), the results do wind up supporting casual deterrence intuitions—i.e., that raising p or s results in less time devoted to crime (Eide 2000, p. 347)—but a great deal always turns on risk attitudes. And, as the models become more realistic, other technical conditions, such as the effect on the illegal gain if the offender is apprehended, become important. Schmidt and Witte (1984, p. 160), for instance, report that even the *source* of a change in p (e.g., an increase in the arrest rate versus an increase in the conviction rate) can result in a stronger or weaker prediction. In their most complicated (read: realistic) deterrence model, Schmidt and Witte (1984, p. 164) acknowledge that unambiguous predictions are not possible without four strong assumptions regarding (1) attitudes toward risk, (2) net returns to crime and the net gains to crime if the individual is convicted, (3) how the marginal utility derived from one activity changes with the time allocated to another, and (4) the way in which the marginal utility of an activity changes with a change in income. They conclude their excellent review (on which Eide (2000) and this entry both draw heavily) by noting:

> Economic models that allow either the level of sanctions or the time allocation to enter the utility function directly are more appealing intuitively. They allow individuals to have different attitudes

> toward sanctions, work, and illegal activity, and they answer the question of how leisure is determined. However, this increased realism is not without its price. These models do not allow us to determine unambiguously how changes in criminal justice practices or opportunities to conduct legal activities will affect participation in illegal activities. The ultimate effect of stiffened penalties [(s)], increased probabilities of apprehension [(p)], or improved legal opportunities becomes an empirical question (Schmidt and Witt 1984, p. 183).

Still, as between increasing criminal sanctions (s) or increasing the degree of enforcement (p) as a means of reducing criminal activity, deterrence models support the relative robustness of increasing p, given the possibility that offenders may be risk loving. As Eide (2000, p. 347) generalizes, "[b]oth the probability and the severity of punishment are found to deter crime for a risk averse person. For risk lovers, the effect of severity of punishment is uncertain." However, if p is disaggregated into the probability of arrest, the probability of conviction conditional on arrest, and so on, different conclusions are possible with respect to the "conditional" levers (Schmidt and Witt 1984, p. 160). It is worth reiterating that other technical assumptions also matter to whether these models can generate useful predictions, such as whether all costs and benefits can be monetized and the precise nature of the enforcement probabilities (Eide 2000, p. 350).

Finally, in most of these models of offending behavior, the level of risk aversion will affect how p and s can be combined to achieve a certain level of deterrence. When risk aversion levels are high, for example, increasing the severity of sanctions will be a comparatively easier strategy for increasing the offender's costs of criminal activity (Shavell 2004, p. 480).

Types of Criminal Sanction: Implications for Deterrence

A rich theoretical literature exists on the relative merits of using monetary fines and nonmonetary sanctions, like incarceration. Many of the key results, however, turn on the social cost of imposing the sanction in question. For instance, all else equal, fines are superior to imprisonment because, theoretically, a fine is less expensive to impose and is effectively just a transfer from the offender to society (i.e., "utility" is destroyed by incarceration, not transferred to someone else). Of course, comprehensive summaries of these results like Shavell (2004) and Polinsky and Shavell (2007) also acknowledge that sanction type is relevant to offender behavior in the first instance and therefore to social welfare calculations, but direct social costs typically feature more prominently. By contrast, this section will focus on a few implications for deterrence of choosing between different types of criminal sanctions.

While attractive on the grounds that they are less socially costly than other types of sanctions, monetary sanctions often face an essential practical difficulty: offenders often, if not usually, have too few assets for the optimal fine to be imposed (Piehl and Williams 2011, p. 117). When this is the case, often with respect to serious crimes, fines will underdeter relative to other sanctions. As the saying goes, "you can't get blood from a stone"; the actual scope of punishment is therefore limited to what can reasonably be extracted from an offender. By definition (or at least the assumption is that), courts and corrections departments are unable to force a person to work (else, the fine becomes nonmonetary). In the limit, if a potential offender has no assets whatsoever and no realistic prospect of acquiring (or incentive to acquire) assets in the future that can be made subject to attachment, fines will produce no deterrence. Furthermore, once assets are exhausted, marginal deterrence is absent, also.

Technically, all sanctions have this feature. For instance, we only have so much time to live, and so a threat of incarceration will fail to deter someone who is certain he will die in the next 24 hours. But there are important reasons why the average potential offender's circumstances are more likely to allow him to avoid (and know he can avoid) a monetary sanction. First, although not an iron law, most potential offenders are financially insecure. Indeed, in the case of crimes with a financial motive, the source of the benefit of the crime may be the offender's lack of assets or alternative options for earning a living. Second, most potential offenders are comparatively young, which means that even if they have no financial resources, they

have other assets—years of their lives—that will serve as a better foundation on which to base a sanction. One can compensate for the limited assets of offenders by increasing the likelihood that any crime is detected, but this is a costly response and is insufficient when adequate deterrence requires a high fine.

Monetary sanctions—or at least significant ones—may also differ in other key ways from nonmonetary sanctions. Most importantly, levels of risk aversion—central in identifying deterrence model predictions—will almost certainly differ depending on whether the potential offender faces fines or nonmonetary sanctions. There is little evidence in support of the idea that potential offenders will be risk loving with respect to an increase in a threatened fine, whereas offenders may anticipate becoming accustomed to long imprisonment after so many years, increasing their relative preference for risk (Shavell 2004, pp. 503 n.17, 508).

Fines also differ from other forms of criminal sanctions in how potential offenders perceive them. Piehl and Williams (2011, p. 116) suggest that fines are viewed as "softer" than other forms of punishment—perhaps just a price, one that is not often paid—and that potential offenders' responses to fines relative to other sanctions will be heterogeneous. Briefly stepping outside of the neoclassical economic framework, cognitive and behavioral biases clearly play an important role in distinguishing types of sanctions in the minds of offenders. Also worth noting is the fact that a monetary fine may be easier to set optimally when the crime is one that results in a financial benefit for the offender or a financial harm to a victim, as the fine imposed can be explicitly linked to these values.

Modern nonmonetary sanctions as a group tend to involve (1) constraints on where an offender can go, what he can do, who he can see, or where he can live (e.g., incarceration, residency restrictions, or banishment), (2) affirmative obligations that require time and effort (e.g., registration requirements), or (3) the imposition of psychic harm, usually through some form of humiliation (e.g., shaming penalties generally or sex offender community notification requirements). Of the possible nonmonetary sanctions, intermediate or alternative sanctions—which were defined above—appear to be increasingly attractive from the government's perspective because they are less costly than incarceration, because they do not require that an offender has wealth or income (Miceli 2009, p. 279), and because technological innovation offers the potential to reclaim otherwise forgone incapacitation benefits. At the same time, there are good reasons to surmise that offenders will perceive these sanctions in very different ways and that the effective cost of these sanctions to the offender will depend on characteristics more difficult to measure than the financial assets or wages of the offender.

Moreover, these sanctions, especially intermediate or alternative sanctions, may generate deterrence using more complicated mechanisms. For instance, relative to fines, we know incarceration may work differently across the board and with respect to specific people: a different asset (opportunities during a period of time) is confiscated from the offender, but money is typically easier to value and to relate to how the offender benefits from committing the crime or the social harm the crime caused. On the whole, though, all sanction types, if set appropriately and with an eye toward these concerns, should be capable of playing the generic role of s, shifting the perceived cost of committing a crime in a way that may cause the offender to use his time differently.

Nevertheless, there are at least two important reasons for explicitly considering the specific roles that the various types of criminal sanctions might play in any model of offending behavior, both of which can add to our understanding of criminal deterrence generally.

First, as pointed out at the end of the discussion of the types of criminal sanctions, the *simultaneous* use of two or more different kinds of sanctions may generate different deterrent effects than a comparable amount of punishment using a single sanction type. (Note: This proposition is distinct from the result from the public enforcement literature that an optimal sanction scheme involves first exhausting wealth through fines before turning to nonmonetary sanctions. This well-known idea turns on fines being superior but limited by the liquidity constraints of the offender.) Prescott and

Rockoff (2011) find evidence consistent with this idea in the sex offender law context, reporting surprisingly strong deterrent effects from the addition of post-release notification requirements (effectively, shaming) to be imposed at the end of already very long sentences. Although comparing the two is complicated, research that has examined simply the addition of more prison time at the end of a long sentence appears to suggest that "more of the same" had weaker effects (Kessler and Levitt 1999).

If there are diseconomies of scale or perhaps economies of scope in the deterrent effects of criminal sanctions, one likely mechanism is that offenders perceive the sanction regimes differently: adding distinct sanctions may seem worse than simply adding more of a sanction already in play. It is also possible that increasing the severity of punishment by adding a new type of sanction does not invite additional risk taking by offenders who are risk lovers, allowing the reduction in criminal activity by the risk averse to more clearly dominate the final deterrence tally.

Second, even if two criminal sanctions are thought to be equivalent in terms of their *deterrent* effects from the perspective of a potential offender, criminal sanctions may differ in how they influence ex post offender behavior down the road, leading potentially to different criminal activity levels in the aggregate. For example, the use of incarceration not only deters potential offenders ex ante but it incapacitates those who do choose to offend, which may offer separate social value in precluding subsequent offenses. There is no a priori reason to think that these indirect effects are always positive nor that they function solely in ways unrelated to deterrence.

Indeed, collateral effects of criminal sanctions may have indirect consequences for deterrence itself—and they may in fact *encourage* offending. Sex offender notification laws, for example, are intended to reproduce an upside of incarceration, essentially incapacitating potential recidivists by using information to aid potential victims in creating a barrier between themselves and released sex offenders. Unfortunately, notification also produces many deprivations in addition to the incapacitation effect: finding employment and housing are difficult, public shaming results in strained relationships, and so on. Furthermore, these deprivations do not depend on whether the offender engages in criminal behavior. As a result, although the threat of notification may deter criminal offending ex ante by raising the costs of committing a sex crime, the application of the requirements ex post may hamper deterrence, by reducing the relative benefits of staying on the straight and narrow (Prescott and Rockoff 2011). Unless the incapacitation effect more than offsets this reduction in deterrence (or if notification increases p sufficiently through additional contact with law enforcement or public monitoring), criminal activity should increase, all else equal.

Empirical Evidence on Deterrence

There is a substantial and growing body of empirical evidence on criminal deterrence, and this entry only offers a brief summary of this research. For recent and detailed discussions of key papers and open questions, see generally Nagin (2013), Durlauf and Nagin (2011), Levitt and Miles (2007), Robinson and Darley (2004), Eide (2000), and Freeman (1999). Recognizing that empirical work must study changes from an existing framework of criminal sanctions and enforcement levels and that there is no reason to think deterrent effects should be homogeneous across the different departure points, much less across all types of crime, types of sanctions, and sanction levels, the work is decidedly mixed. Notwithstanding these caveats and all of the others implicit in empirical work in general, some basic points of conventional wisdom have emerged.

First, as a general matter, most scholars interpret the literature as strongly supporting the ability of criminal sanctions and enforcement to deter criminal behavior. The evidence in favor of "substantial" deterrent effects across "a range of contexts," according to some, is "overwhelming" (Durlauf and Nagin 2011, p. 43). The details, however, reveal quite a bit of variation (in magnitudes, in particular), and others strongly disagree with this general conclusion—at least in

terms of the policy implication that an increase in *s* or *p* from present levels would result in less crime—on the basis of alternative evidence that appears inconsistent with it (like evidence that suggests many offenders are unaware of the size of criminal sanctions) and methodological objections (see, e.g., Robinson and Darley 2004), which are particularly strong with respect to early work and which are discussed in most reviews. Eide (2000, pp. 364–368) also highlights the importance of properly interpreting the empirical evidence.

Second, most scholars agree that increasing the degree of enforcement (*p*) (often referred to as "certainty" in the literature) appears to be more effective than a comparable increase in the severity of sanctions, at least given the current levels of each. Consistent with the models of Schmidt and Witte (1984, p. 160), Nagin (2013, p. 201) concludes that the evidence in favor of certainty is much stronger with respect to the probability of arrest and less so (or at least with more uncertainty) with respect to conviction and other later conditional enforcement probabilities. To the extent that deterrence is an important goal of criminal law, the take-away from these patterns is that crime policy should focus on how laws are enforced rather than on the severity of sanctions.

Third, all agree not only that increasing the severity of sanctions appears less effective than increasing the degree of enforcement but also that there is still some question whether, at current sanction levels, increasing severity will result in *any* reduction in crime (Durlauf and Nagin 2011). Studies of the effects of increased severity on crime are much more likely to produce estimates that are not statistically significant, and at times, the estimates are even positive (Eide 2000, p. 360). This body of work is actually remarkably consistent with what one would have expected to find given the conclusions of the theoretical work. In the standard models of criminal behavior, increasing sanction severity, for instance, was more likely to lead to more crime all else equal than increasing the certainty of enforcement, especially to the extent that the offenders in question are risk loving.

Concluding Remarks

The goal of this entry has been to contextualize the law and economics of criminal deterrence by linking research in economics to the range of criminal sanctions that exist in the real world. Some emphasis has been placed on the role intermediate or alternative sanctions might play in the deterrence framework, as there appears to be very little theoretical work on the subject. Unfortunately, many basic ideas from the research on deterrence are not reviewed in this entry. Optimal enforcement policy has only been mentioned in passing, and many key ideas in the deterrence literature have been omitted: marginal deterrence (see Shavell 2004, p. 518), specific deterrence (see Shavell 2004, p. 516), the roles that can be played by a fault standard or affirmative defenses in designing sanctions (see Shavell 2004, pp. 476, 494–495), and how best to deter repeat offenders (see Polinsky and Shavell 2007, p. 438).

References

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217

Bierschbach RA, Stein A (2005) Overenforcement. Georgetown Law J 93:1743–1781

Calabresi G, Douglas Melamed A (1972) Property rules, liability rules, and inalienability: one view of the cathedral. Harv Law Rev 85:1089–1128

Durlauf SN, Nagin DS (2011) The deterrent effect of imprisonment. In: Cook PJ, Ludwig J, McCrary J (eds) Controlling crime: strategies and tradeoffs. University of Chicago Press, Chicago, pp 43–94

Eide E (2000) Economics of criminal behavior. In: Boudewijn B, De Gerrit G (eds) Encyclopedia of law and economics, vol 5. Edward Elgar, Cheltenham, pp 345–389

Freeman RB (1999) Chapter 52: The economics of crime. In: Ashenfelter O, Card D (eds) Handbook of labor economics, vol 3. North-Holland, Amsterdam, pp 3529–3571

Kessler DP, Levitt SD (1999) Using sentence enhancements to distinguish between deterrence and incapacitation. J Law Econ 42(1):343–363

Levitt SD, Miles TJ (2007) Chapter 7: Empirical study of criminal punishment. In: Polinsky AM, Shavell S (eds) Handbook of law and economics, vol 1. North-Holland, Amsterdam, pp 457–495

Miceli T (2009) The economic approach to law, 2nd edn. Stanford Economics and Finance, Stanford

Nagin DS (2013) "Deterrence in the twenty-first century". Crime and justice. Crime Justice Am 1975–2025 42(1):199–263
Piehl AM, Williams G (2011) Institutional requirements for effective imposition of fines. In: Cook PJ, Ludwig J, McCrary J (eds) Controlling crime: strategies and tradeoffs. University of Chicago Press, Chicago, pp 95–121
Polinsky AM, Shavell S (2007) Chapter 6: The theory of public enforcement of the law. In: Polinsky AM, Shavell S (eds) Handbook of law and economics, vol 1. North-Holland, Amsterdam, pp 457–495
Prescott JJ, Rockoff JE (2011) Do sex offender registration and notification laws affect criminal behavior? J Law Econ 54(1):161–206
Robinson PH, Darley JM (2003) The role of deterrence in the formulation of criminal law rules: at its worst when doing its best. Georgetown Law J 91: 949–1002
Robinson PH, Darley JM (2004) Does criminal law deter? A behavioral science investigation. Oxf J Leg Stud 23(2):173–205
Schmidt P, Witte AD (1984) Chapter 9: Economic models of criminal behavior. In: An economic analysis of crime and justice: theory, methods, and applications. Academic, Orlando, pp 142–193
Shavell S (2004) Foundations of economic analysis of law. Belknap Harvard.

Currency Demand

► Cash Demand

Custodian Liability

Jef De Mot[1] and Louis Visscher[2]
[1]Postdoctoral Researcher FWO, Faculty of Law, University of Ghent, Belgium
[2]Rotterdam Institute of Law and Economics (RILE), Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Definition

Liability of a person for losses caused by an object under that person's control.

The Law on Custodian Liability

Custodian liability means that a person is liable for losses caused by an object under that person's control. It has its origins in the French law (Wagner 2012). Article 1384(1) of the French Code Civil states that a person is responsible for the damage caused by things in his custody. Originally, this phrase was not intended as a separate cause of action but merely as a reference to the provisions 1385 and 1386 CC, which respectively hold keepers of animals and owners of buildings liable for damage caused by animals or defective buildings. Some landmark judgments by the Cour de Cassation however transformed this statement into the legal basis for making custodians liable for damages caused by objects of *any kind*. A claimant needs to show that the thing has contributed to the realization of the damage. A thing includes all inanimate objects. It can be dangerous or not dangerous, movable or immovable (e.g., a lift, a dyke), natural or artificial, and defective or nondefective. The contribution of the thing can be established in several ways (Fabre-Magnan 2010). If there has been contact between the claimant or his property and a *moving* thing (e.g., a falling tree or a rolling stone), it is simply assumed that the thing has contributed to the realization of the damage. The claimant does not need to show that the custodian contributed to the action of the thing. If there was contact between the claimant or his property and a *nonmoving* thing, the claimant needs to prove that the thing was instrumental in causing the damage in some way. He can do this by showing that the thing behaved in an abnormal way, more particularly that the thing was defective or was in an abnormal position. Finally, if there was no contact between the claimant or his property and the thing, Art. 1384(1) may apply, but the claimant needs to prove that the thing was the cause of his damage. The liable person in Article 1384(1) is the *custodian*: the person who had, at the time of the accident, the power to use, control, and direct the thing. Often this will be the owner of the thing, but also people other than the owner can be the custodian. The custodian can escape liability if he can prove an external cause of the damage which was both unforeseeable and unavoidable. The external cause

can be due to *force majeure*, the act of a third part or the act of the victim. This defense is generally rather difficult to prove. Also, the custodian can invoke the claimant's contributory negligence. If successful, the amount of compensation the custodian needs to pay is reduced.

Many legal systems that have adopted or integrated the French Code Civil have been influenced by the French model of custodian liability (Wagner 2012). The concept of custodian liability is known in, for example, Italy, Portugal, the Netherlands, Belgium, and Luxembourg. One important exception is Spain. Custodian liability is also unknown in German law, English common law, and more generally in countries outside the gravitational force of the Code Civil. Some countries which have introduced custodian liability limit this concept to cases where the object suffers from a defect (e.g., the Netherlands and Belgium). It is a general requirement that the object in question represents a source of abnormal danger to its surroundings. Case law in this respect frequently refers to the normal expectations in society with respect to the characteristics of the object.

The remainder of this contribution is structured as follows. Section "Advantages of Strict Liability and Negligence in the Context of Liability for Harm Caused by Objects" briefly discusses the relative advantages of strict liability and negligence for damage caused by objects under a person's control. Section "Economic Analysis of Case Law on Custodian Liability" provides an overview of how some European countries have actually applied the several concepts of custodian liability in an economically efficient way. Section "Conclusion" concludes.

Advantages of Strict Liability and Negligence in the Context of Liability for Harm Caused by Objects

Should custodian liability be an application of strict liability or negligence? In this section, we provide a brief overview of the main relative advantages of both rules (also see Schäfer and Müller-Langer 2009, pp. 3–45). There can be sound economic reasons to hold the custodian strictly liable for losses caused by an object under his control:

- It provides incentives to the liable party to maintain his things, to regularly check and if necessary repair them, to replace components, etc.. In theory, negligence could also provide those incentives, but then the victim has to prove that the custodian did not take enough care and that he should have known about the existence of a defect. This will often be very difficult since many care measures may be non-verifiable (e.g., how often did the injurer check the brakes of his bicycle). Negligence would therefore frequently not result in liability and the behavioral incentives of liability would be frustrated. Therefore, the superior information of the custodian regarding which care measures he actually *has taken* pleads for strict liability. Especially with respect to immovables, the custodian will generally be better informed about the quality of the building or structure and about the care measures that are taken and that can be taken.
- Negligence can only give care incentives to the injurer in the dimensions that are incorporated in the negligence rule. Strict liability gives the injurer care incentives also in non-verifiable dimensions, which cannot be incorporated in a behavioral norm. On the other hand, when it is more important to provide care incentives in all dimensions to the victim rather than to the injurer, negligence is to be preferred: only then will the victim (being the residual loss bearer) receive care incentives in all dimensions.
- Similarly, strict liability is preferable when it is more important to control the activity level of the custodian than that of the victim (Shavell 1980, p. 2; Landes and Posner 1987, p. 61; Faure 2002, p. 366; Posner 2003, pp. 177, 178; Shavell 2004, p. 193). Many accidents can be regarded as unilateral because the victim has no influence on their incidence. In bilateral cases, where also the victim should take care, a defense of contributory or comparative negligence is necessary under strict liability to avoid that the victim takes no care at all because the injurer would be liable anyway.

- If the injurer has better information than the courts regarding which care measures he *can* take and what the costs and benefits of such measures are, strict liability is to be preferred. There the injurer himself weighs costs and benefits of care, whereas under negligence the court has to do that (Shavell 2004, pp. 18, 188; Landes and Posner 1987, p. 65).
- According to many law and economics scholars, the problem of judgment proofness is more severe under strict liability because the "reward" for being careful (reducing the accident probability) is lower than under negligence (fully escaping liability). Taking care under negligence is therefore also worthwhile for lower levels of wealth.
- If the injurer is a firm, strict liability induces loss spreading via the price of the product/service, whereas negligence leaves the loss (concentrated) with the victim.
- Insurance issues such as moral hazard and adverse selection often give a preference for negligence. Especially in situations which could be regarded as unilateral, an insured victim cannot display moral hazard, whereas an insured injurer could (Bishop 1983, p. 260; Faure 2005, p. 245, 261). In as far as liability insurance makes use of experience-rated premiums, deductibles, etc., the relative importance of this issue may be limited.
- Finally, administrative costs are generally assumed to be lower under strict liability, especially because more settlements are expected (there is no dispute over the true and due levels of care).

Economic Analysis of Case Law on Custodian Liability

De Mot and Visscher (2014) argue that Dutch and Belgian courts apply custodian liability in an economically efficient way (full references to the cases mentioned in this chapter can be found in De Mot and Visscher (2014)). First, the courts use the concept of defect to allocate costs to the least cost avoider. Courts often conclude that there is a defect when it is more important to provide incentives to the injurer to take (often) unverifiable precautions and that there is no defect when it is more important to provide incentives to the victim to take such precautions. This way, the accident losses are allocated to the "least cost avoider." This is especially clear in cases where the floor of a shop becomes slippery (or more dangerous) due to something laying on the floor, e.g., a vegetable leaf in the vegetable department of a supermarket. According to Belgian and Dutch courts, the mere presence of the leaf does not make the building defective. In a Dutch case, the court first noted that the floor is cleaned every morning, that there is always personnel present which is instructed to pick up litter if they observe it, and that there is active monitoring on littering. These measures are relatively easy to verify. What is much more difficult to verify is (1) whether the personnel always picks up litter when they observe it and (2) how careful the client walks in the shop. So the question becomes: if we can only give one party incentives to engage in non-verifiable measures, who do we choose to incentivize? Which party's non-verifiable measures are most productive? Given that it is unreasonable (read: too costly) to require the shop to keep the floor clean all the time, clients have to realize that vegetable leaves (or comparable litter) may lie on the floor. It is thus important that clients are prudent. They are the least cost avoiders (with respect to the non-verifiable precautions). A comparable decision has been taken in Belgian and Dutch cases where someone slipped on the floor of a supermarket near the entrance which was wet due to the rainy weather and soaked carpets. Again, it is more difficult for the supermarket to avoid the floor from becoming wet and slippery also in situations of rain than it is for individuals to walk carefully in situations where the water and dirt were clearly visible. As the Belgian court stated: "During rainy weather, it is impossible for a supermarket to make sure the floor is dry continuously." Furthermore, in cases in which the courts decided that there actually *was* a defect, it is clear that the customer was *not* the least cost avoider. The courts, for example, decided that the supermarket floor was defective in the following cases: ice cream on the floor of

the perfume department, oil on the floor of the food department, and a vegetable leaf on the floor in the textile department. Customers do not expect ice cream in the perfume department or vegetable leaves in the textile department, and they cannot be the least cost avoider. The shop owner is better placed to instruct his personnel to be watchful. As the court stated in the case of oil on the floor of the food department: "the (lower) court could lawfully decide that the floor of the food department was defective by establishing that the floor was exceptionally slippery due to an oil stain which was difficult to observe for a normally attentive customer and that an oil stain in this department is not an everyday phenomenon which clients reasonably have to expect and avoid." De Mot and Visscher (2014) discuss many other slip and fall cases (e.g., a woman tripping over a gasoline hose at a gasoline station, a customer slipping over French fries in a chips shop, a bicyclist slipping when crossing wet train tracks) and other types of cases (e.g., a man ordering pheasant in a restaurant and damaging his teeth because there was still some lead in the pheasant, a shed catching fire while welding work was executed there) in which the courts make their decision in an economically efficient way.

Second, in some cases the discussion is not about whether the object is defective, but there is disagreement over who is the custodian. In these cases, the courts fill in the *concept of custodian* as if they place liability on the least cost avoider. For example, a player of a sports club got injured due to an iron bar sticking out from the field but hidden under the grass. The sports club had permission from the renter of the field to use the field for a couple of hours. The Belgian court concluded that the renter and not the owner of the field or the sports club was the custodian of the sports field, since the renter was the party who "called the shots" and took all the factual decisions with respect to the field; the sports club was only permitted to use the field after permission of the renter and only for a couple of hours and in the condition that the field was in. Hence, the renter was better placed to prevent the accident than the owner or the sports club. In a Dutch case, the victim fell off a stairway to the basement. He was an employee of a telecom company who had to repair a connection in a shop, which was located at the top of a stairway. The employee walked down the stairway because the lid of the connection point was missing, and he wanted to check if it had fallen onto the basement floor. The fore-last step of the stairway turned out to be missing and the employee fell down the stairway. The court holds the renter of the building, who operated his shop there, strictly liable because there was a step missing and there was no adequate warning against this danger. The renter argued that the owner of the building should be held liable, but the court applies Art. 6:181 CC, which states that the business user rather than the possessor is the liable party. This makes economic sense: the renter of a building uses the building on a daily basis and is better informed about possible risks (such as a missing step) than the owner, who might not enter his rented building for substantial periods of time. It is therefore better to incentivize the renter to use his superior information (either by repairing the problem himself or by informing the owner that he should repair the stairway) than to require the owner to regularly check his buildings in which another party operates his business. More generally, in cases in which it is debated whether the letter or the renter was the custodian, Belgian courts have taken into account some elements that can obviously be linked to the least-cost-avoider concept. The following elements make it more likely that the letter is the custodian and not the renter (see Vansweevelt and Weyts (2009), pp. 485–487): limits on the use of the object for the renter (the object can only be used for a limited period, for a specific goal, or in a certain area), whether the letter is present when the object is used and whether the letter is responsible for the maintenance of the object. These considerations make economic sense. When one or more of these conditions is fulfilled, it is more likely that the letter and not the renter is the least cost avoider.

Conclusion

Economic analysis of tort law has shown under which circumstances strict liability creates better

incentives than negligence and vice versa. Often, whole classes of cases in which strict liability should apply can be described in a rather simple way (e.g., for transporting and/or working with dangerous substances). In other types of cases, however, like slip and fall cases, what is the optimal rule is more ambiguous. In some cases, strict liability leads to better results, and in other cases negligence is superior. In some European countries, this difficult exercise is performed by the courts through the interpretation of the concepts defect and custodian. The variety of solutions which is reached in this manner cannot be fully replicated with a binary choice between strict liability with a defense of contributory negligence and negligence with a defense of contributory negligence. If we choose a rule of strict liability with a defense of contributory negligence, we cannot reach optimal solutions in cases in which the victim's non-verifiable measures are very likely to be the most productive ones. And if we choose a rule of negligence with a defense of contributory negligence, we cannot reach optimal solutions in cases in which the injurer's non-verifiable measures are very likely to be the most productive ones. The interesting feature of the concept of a defect as it is interpreted by the Belgian and Dutch courts is that it enables the court to select the liability rule that is most suitable for the case at hand.

This field of tort law can still benefit from further economic research. For example, De Mot and Visscher (2014) have focused on Dutch and Belgian case law. Whether courts in other jurisdictions interpret the various concepts related to custodian liability in an economically efficient way is yet to be discovered. Also, it would be interesting to compare the concepts of defect and custodian with the concepts that are used in countries in which custodian liability is unknown and which may also have the effect of efficiently sorting between cases.

References

Bishop W (1983) The contract-tort boundary and the economics of insurance. J Legal Stud 12:241–266

De Mot J, Visscher LT (2014) Efficient court decisions and limiting insurers' right of recourse. The case of custodian liability in the Netherlands and Belgium. Geneva Papers Risk Insur – Issues Practice forthcoming, 39 (3):527–544

Fabre-Magnan M (2010) Responsabilité Civile et Quasi-Contrats, 2nd edn. Presses Universitaires de France, Paris

Faure MG (2002) Economic analysis. In: Koch BA, Koziol H (eds) Unification of tort law: strict liability. Kluwer Law International, Den Haag, pp 361–394

Faure MG (2005) The view from law and economics. In: Wagner G (ed) Tort law and liability insurance. Springer, Vienna, pp 239–273

Landes WM, Posner RA (1987) The economic structure of tort law. Harvard University Press, Cambridge, MA

Posner RA (2003) Economic analysis of law, 6th edn. Aspen Publishers, New York

Schäfer H-B, Müller-Langer F (2009) Strict liability versus negligence. In: Faure M (ed) Tort law and economics, vol 1, 2nd edn, Encyclopedia of law and economics. Edward Elgar, Cheltenham, pp 3–45

Shavell S (1980) Strict liability versus negligence. J Legal Stud 9:1–25

Shavell S (2004) Foundations of economic analysis of law. The Belknap Press of Harvard University Press, Cambridge, MA

Vansweevelt T, Weyts B (2009) Handboek Buitencontractueel Aansprakelijkheidsrecht. Intersentia, Antwerp

Wagner G (2012) Custodians liability. In: Basedow J, Hopt KJ, Zimmerman R, Stier A (eds) Max Planck encyclopedia of European private law. Oxford University Press, Oxford

Custom of Merchants

▶ Lex Mercatoria

Customary Law

Bruce L. Benson
Department of Economics, Florida State University, Tallahassee, FL, USA

Abstract

Customary law, a system of rules of obligation and governance processes that spontaneously evolve from the bottom up within a community, guides behavior in primitive, medieval,

and contemporary tribal societies, as well as merchant communities during the high middle ages, modern international trade, and many other historical and current settings. Rules and procedures are recognized and accepted because of trust arrangements, reciprocities, mutual insurance, and reputation mechanisms, including ostracism threats. Negotiation (contracting) generally is the most important source for initiating change in customary law. Such agreements only apply to the parties involved, but others can voluntarily adopt the change if it proves to be beneficial. An individual also may unilaterally adopt behavior that others observe, come to expect, and emulate, or a dispute may arise that results in an innovative solution offered by a third party and voluntarily adopted by others. Many different third-party dispute resolutions procedures are observed in different customary communities, but they generally involve experienced mediators or arbitrators who are highly regarded community members. Some of these adjudicators may have leadership status, but leadership arises through persuasion and leaders do not have coercive authority. Since there is no coercive authority, protection and policing rely on voluntary arrangements, and community norms encourage and reward such activity. People who violate rules are generally expected to and have incentives to compensate victims for harms. Customary law is polycentric, with hierarchical arrangements to deal with intercommunity interactions. Customary law also may conflict with authoritarian law. When this occurs, a coercive authority may attempt to assert jurisdiction over a customary-law community, but this will have very different impacts depending on the options available to members of the community. The authority often adopts and enforces some customary rules in order to avoid conflict.

Synonyms

Informal law, Polycentric law, Private law, Stateless law

Definition

Customary law: a system of rules of obligation and governance processes that spontaneously evolve from the bottom up within a community.

"Customary law" sometimes refers to various immanent principles, so well established and widely recognized that sovereigns feel obliged to adopt them as law (e.g., as in common law recognition of "immemorial custom"), but it also can be applied to the legal arrangements within primitive communities. Another definition delineates the source of immanent customary principles, however, and encompasses primitive law: customary law is a system of rules of obligation and governance processes that spontaneously evolve from the bottom up within a community (Pospisil 1971). This definition is the focus of the following presentation. Those who contend that "law" consists of general commands of a sovereign will not consider this concept to be appropriately labeled as law, nor will those who contend that "law" applies to moral principles, whether logically derived or handed down by a higher being. Nonetheless, these rules and processes are real, and they influence behavior in very significant ways. This is true for the primitive (Pospisil 1971; Benson 1991), medieval (Friedman 1979), and contemporary tribal societies (Benson and Siddiqui 2014), merchant communities during the high middle ages (Benson 1989, 2014), modern international trade (Benson 1989), and many other historical and current settings (Fuller 1981).

Kreps (1990) stresses that socially or culturally derived experiences help players "know" what to do and predict what other players will do. In this context, Hayek (1973, pp. 96–97) observes that many issues of "law" are not "whether the parties have abused anybody's will, but whether their actions have conformed to expectations which other parties had reasonably formed because they corresponded to the practices on which the everyday conduct of the members of the group was based. The significance of customs here is that they give rise to expectations that guide people's actions, and what will be regarded as binding will therefore be those practices that everybody counts on being observed and which thereby

condition the success of most activities." Similarly, Fuller (1981, p. 213) explains that "We sometimes speak of customary law as offering an unwritten code of conduct. The word *code* is appropriate here because what is involved is not simply a negation, ... but of this negation, the meaning it confers on foreseeable and approved actions, which then furnish a point of orientation for ongoing interactive responses." In this light, behavioral patterns individuals within a community are generally expected to adopt and follow in their various interdependent activities are considered to be legal rules below, whether those rules (expectations) arise through formal legislation or informal customary processes discussed below. Individuals are expected to follow such rules, and these expectations influence the choices made by other individuals. Customary law involves more than rules of behavior, however, as customary governance processes also evolve from the bottom up (Benson 1989, 2011, 2014; Pospisil 1971). Development and characteristics of both customary behavioral rules and procedural rules are discussed below.

Establishing and Recognizing Customary Rules

Vanberg and Buchanan (1990, p. 18) define rules of behavior toward others which individuals have positive incentives to voluntarily recognize as "trust rules" and explain that:

> By his compliance or non-compliance with trust rules, a person selectively affects specific other persons. Because compliance and non-compliance with trust rules are thus "targeted," the possibility exists of forming cooperative clusters.... Even in an otherwise totally dishonest world, any two individuals who start to deal with each other - by keeping promises, respecting property, and so on - would fare better than their fellows because of the gains from cooperation that they would be able to realize.

Game theory demonstrates that such cooperation can arise through repeated interactions, although the dominant strategy in bilateral games still depends on expected payoffs, frequency of interaction, time horizons, and other considerations (Ridley 1996, pp. 74–75). As North (1990, p. 15) explains, however, game theory "does not provide us with a theory of the underlying costs of transacting and how those costs are altered by different institutional structures." For example, if bilateral relationships form in recognition of the benefits from cooperation in repeated games, and if individuals in such relationships enter into similar arrangements with other individuals, a loose knit group with intermeshing reciprocal relationships develops. As this occurs, tit for tat becomes less significant as a threat, while expected payoffs from adopting trust rules increase. For instance, an exit threat becomes credible when each individual is involved in several different games with different players, in part because the same benefits of cooperation may be available from alternative sources (Vanberg and Congelton 1992, p. 426).

When the exit option becomes viable, Vanberg and Congleton (1992, p. 421) explain that a potential strategy is unconditional cooperation unless uncooperative behavior is confronted, and then imposition of some form of explicit punishment on the noncooperative player as exit occurs. They label such a strategy "retributive morality"; examples include the "blood feuds" of tribal (Benson and Siddiqui 2014) and medieval (Friedman 1979) societies. Retributive morality strengthens the threat against noncooperative behavior relative to tit for tat. Such violence is risky, however, and there is likely to be a better alternative. Since all community members have exit options, information about uncooperative behavior can be spread, creating incentives for everyone to avoid interacting with the untrustworthy individual. This suggests a strategy involving unconditional cooperation in all interactions with other community members, along with a refusal to interact with any individual who is known to have adopted noncooperative behavior with anyone in the group and the spread of information about such behavior. Vanberg and Congleton (1992) refer to this response as "prudent morality," and given that reputation information spreads quickly and everyone spontaneously responds to information, the result is spontaneous social ostracism. Depending in part on severity of the offenses, however, ostracism may not be

absolute. Individuals might continue interacting with someone who has misbehaved, but only if certain conditions are met to reduce risk (e.g., a bond might have to be posted). Such conditions sanction an offender by raising costs or reducing his benefits in various interactions.

Many group-wide customary rules are simply commonly shared trust rules. Others, called "solidarity rules" by Vanberg and Buchanan (1990, pp. 185–186), are expected to be followed by all members of the group. The spontaneous development of social ostracism illustrates this. Solidarity rules produce community-wide benefits (Vanberg and Buchanan 1990, p. 115), including general deterrence and others discussed below.

Changing Customary Rules

Rules and communities evolve simultaneously: the evolution of trust rules leads to the development of a web of interrelationships that become a "close-knit" community, and the evolving web of interactions and expanding opportunities for interactions in turn facilitates the evolution of more rules. If conditions change or a new opportunity is recognized, for instance, and a set of individuals decide that, for their purposes, a new behavioral obligation arrangement will support more mutual benefits, they can voluntarily agree to accept (contract to adopt) the obligations. Such new obligations only apply for the contracting parties for the term of the contract. Individuals who interact with these parties learn about their contractual innovation, however, and/or members of the community observe its results. If the results are desirable, the new rules can be rapidly emulated. Such changes are initiated without prior consent of or simultaneous recognition by any other members of the relevant community, but they can be spread through voluntary adoption. Indeed, as Fuller (1981, pp. 224–225) explains, "If we permit ourselves to think of contract law as the 'law' that parties themselves bring into existence by their agreement, the transition from customary law to contract law becomes a very easy one indeed." In fact, contract and custom are tightly intertwined and often inseparable:

> if problems arise which are left without verbal solution in the parties' contract these will commonly be resolved by asking what "standard practice" is with respect to the issues. . . In such a case it is difficult to know whether to say that . . . the parties became subject to a governing body of customary law or to say that they have by tacit agreement incorporated standard practice into the terms of the contract.
>
> . . . [Furthermore,] . . . the parties may have conducted themselves toward one another in such a way that one can say that a tacit exchange of promises has taken place. Here the analogy between contract and customary law approaches identity. (Fuller 1981, p. 176)

Negotiation (contracting) generally is the most important source for initiating change in customary law (Fuller 1981, p. 157), although there are others. An individual may unilaterally adopt behavior that others observe, come to expect, and emulate, for instance, or a dispute may arise that results in an innovative solution offered by a third party.

If new obligations are required to deal with the new situation, transaction costs may prevent individuals from agreeing on an appropriate arrangement. Similarly, one party may believe that a particular rule applies to a situation, while another may believe that a different rule is relevant. If direct negotiation fails, a solution may still be achieved if the parties turn to arbitration or mediation. Many different third-party dispute resolution procedures are observed in different customary communities (Benson 1991, 2011, 2014), but they generally involve experienced mediators or arbitrators who are highly regarded community members. They may be paid or they may voluntarily give their time in order to enhance their reputations. Whatever the process, a potential rule may be suggested by the dispute resolution (Fuller 1981, pp. 110–111). Unlike modern common law precedent, however, the resolution only applies to the parties in the dispute. If it suggests behavior that effectively facilitates desirable interactions, the implied rule can be adopted and spread through the community.

Reciprocity and Restitution

When retributive morality dominates, unilateral exaction of punishment can be very risky, as

noted above, and generate negative spillovers for the larger community. As a result, customary rules tend to evolve to reduce revenge seeking and encourage substitution of ostracism for retribution, as suggested above, but another option also commonly arises. To understand why, note that unregulated retaliation may result in greater costs for the alleged offender than the costs generated by the offender's initial offense, leading to an escalating chain of violence. In a dynamic society, this ongoing feud consumes resources, including human life, and, therefore, dissipates wealth. In subsistence societies the loss of such wealth can be devastating, but even as wealth expands, the opportunity cost of escalating feuds is high. Some of these costs are also likely to be external. Therefore, strong incentives arise among community members to constrain retribution. Parisi (2001) explains that rules generally evolve to specify who can pursue retribution (e.g., only the victim or a member of his extended family or support group discussed below) and to set an upper bound on the harm imposed on the offender based on the harm initially inflicted on the victim. Over time, the level of retribution moves to one of symmetry, "an eye for an eye." Once the severity of punishment is generally expected to be proportional, it becomes predictable, and the transaction costs of bargaining fall because the parties can bargain over a known "commodity." If the offender is willing to pay compensation that at least offsets the value the victim places on revenge, then violence can be avoided. In fact, in many customary-law systems, offenders reestablish membership in the community when appropriate payments are made. So-called blood money becomes increasingly prevalent.

History and anthropology suggest that customary restitution rules may be quite simple or very complex. For example, the *Hebrew Bible* dictates that "When anyone, man or woman, wrongs another. . ., that person has incurred guilt which demands reparation. He shall confess the sin he has committed, make restitution in full with the addition of one fifth, and give it to the man to whom compensation is due" (*The New English Bible*, Numbers 5: 6–7). Some customary legal systems include widely recognized rules detailing payments for virtually every type of predictable harm (e.g., Goldsmidt 1951; Benson 1991, 2011; Barton 1967). In some societies, including medieval Iceland (Friedman 1979), the payment also depends in part on whether the offender tries to hide or deny the offense. If the offender admits guilt, thereby lowering the costs of enforcement, the payment is lower. Repeat offenders are also treated differently in many restitution-based systems. In Anglo-Saxon England, for instance, an offender could "buy back the peace" on a first offense, but for some kinds of illegal acts, a second offense was not be forgiven. Such an offender becomes an outlaw with no protection, making him fair game for anyone who wanted to attack him. Restitution-based systems also account for the problem of collections from potentially "judgment-proof" offenders. Payments do not necessarily have to be monetary, for instance, as labor services or other "goods" can serve as restitution. In Anglo-Saxon England offenders had up to a year to pay large awards, and if more time was required, they become "indentured servants" until the debt was worked off (Benson 2011, p. 25; 1994).

Bargaining power may differ between individuals, leading to variance in restitution for similar harms, so not surprisingly reciprocal mutual support groups typically develop in customary-law communities. These groups may consist of family members, for instance, but they also may be based on neighborhoods, religious affiliation, ethnicity, participation in the same commercial activity, or some other factor. Such groups accept reciprocal obligations to assist each other in the pursuit of justice, although they generally have many other reciprocal expectations, including social, religious, and joint production activities. Mutual support groups also may pay the restitution to a victim for someone in the group who cannot pay immediately, so the offender is obliged to pay members of his own support group. When such groups develop surety obligations they also may purchase indentured-servitude contracts so offenders work for members of their own groups rather than for their victims.

Clearly, hardships imposed on wealthy offenders who pay a fixed restitution are less

significant than the same payment for poor offenders. Thus, restitution may require a relatively large payment by a wealthy offender, enhancing their deterrence incentives. On the other hand, if a person is diverted from earning income to pursue and prosecute an offender, the value of lost time will be much higher for an individual earning a high income than for a low-income person. Therefore, restitution payments for the same offenses may be higher for the well-to-do victim in order to induce participation in the legal process. The schedule of payments in the *wergeld* or "man price" systems in Anglo-Saxon England reflects the status of the parties involved, as do some tribal customary-law systems (Barton 1967).

Negotiation between a victim and offender may be very difficult, if not impossible, of course, without additional options. As a consequence, a customary rule often develops which requires guilty parties to express remorse or repentance, thereby reducing the costs of negotiation. Substitutes for negotiation are also attractive. A community leader or group of leaders may come forward to encourage repentance and a truce so that negotiations can occur (Benson and Siddiqui 2014). Given the potential animosity between the parties, this third party might also mediate or arbitrate the dispute, or specialists in mediation or arbitration may be called upon. The primacy of rights also changes over time, as a right to restitution supersedes the right to retaliation. In early medieval England and Iceland, and in the large number of tribal societies, victims do not have the right to exact physical punish (retributive morality) unless and until the offender refuses to accept the customary adjudication procedures and/or to pay fair restitution (Benson 2011, pp. 11–30, 1991; Friedman 1979; Pospisil 1971; Goldsmidt 1951; Barton 1967). Similarly, ostracism occurs only when an offender refuses to adjudicate or to pay restitution.

Mutual Insurance

In part, to encourage people to continue to recognize customary behavioral rules, customary-law communities develop mutual insurance arrangements to aid individuals who find themselves in significant risk as a consequence of mistakes, unanticipated natural disasters, warfare, theft, or general bad luck. Johnsen's (1986) analysis of the potlatch system of the Southern Kwakiutl Indians provides an insightful example. He explains that "In order to provide the incentives of would-be encroachers to recognize exclusive property rights, and thus to prevent violence, those Kwakiutl kinship groups whose fishing seasons were relatively successful transferred wealth through the potlatch system to those groups whose seasons were not successful.... Although potlatching thereby served as a form of insurance, the relevant constraint in its adoption and survival was the cost of enforcing exclusive property rights rather than simple risk aversion" (Johnsen 1986, p. 42). Sharing norms such as hospitality and potlatching are common practices in customary-law societies all over the world (Ridley 1996, pp. 114–124). Therefore, even those who may find themselves in desperate situations know that recovery is possible, so they have relatively strong incentives to live up to customary obligations.

Leadership

Customary-law communities do not have executives with coercive power to induce recognition of law. Leaders arise to facilitate various kinds of cooperation, but they lead through persuasion and example. Leadership also is conditional, with no specified terms or binding claims to loyalty. Leaders generally serve as a nexus of the voluntary relationships that dominated the internal life of a community. Therefore, a very important leadership characteristic is a reputation for making good decisions (wisdom) that benefit his followers. Anyone who acquires a leadership role is likely to be a mature, skilled individual with considerable physical ability and intellectual experience and perhaps, more importantly, someone who has a history of cooperative behavior. The importance of maturity and experience often mean that leaders are relatively old community members (elders). They earn respect (Benson

1991; Pospisil 1971), but they also must continue to earn it. Individuals who achieve leadership positions but then make decisions that turn out to be undesirable to followers lose those followers. Wisdom and respect also are not sufficient to attract substantial numbers of followers (Pospisil 1971, p. 67). A self-interested entrepreneurial leader within a close-knit community rationally chooses to pursue activities that benefited others and/or generously spreads the wealth he gains. "The way in which [leadership or social] capital is acquired and how it is used make a great difference; the [members of the community] . . . favor rich candidates who are generous and honest" (Pospisil 1971, p. 67). Indeed, in customary-law societies, the honor of being recognized as a leader is often "purchased" through repeated public displays of generosity demonstrated at occasions such as marriages. Third-party mediation or arbitration often develops in close-knit groups, as noted above, and individuals seeking recognition as leaders also may offer to help resolve disputes and provide mediation or arbitration advice free of charge (Pospisil 1971; Benson 1991). "Fair" nonviolent dispute resolution is attractive to community members because it avoids the spillover costs associated with violent dispute resolution, and it is attractive to the "suppliers" of such dispute resolution even when they are not explicitly paid, because it enhances their prestige. In other words, both generous gift and advice giving (e.g., dispute resolution) signal that the individual is wise, successful, cooperative, and trustworthy. As Ridley (1996, p. 138) puts it, such acts "scream out 'I am an altruist; trust me.'" Leaders engage in such displays of generosity because they expect to benefit in the future through reciprocal obligations and cooperation in joint production. Since leadership positions are available to anyone who can persuade a group to accept his decisions and guidance, customary communities often have multiple leaders. Competition for followers arises, and community members may change their primary allegiances if they feel that a leader has failed to perform well or that another individual offers what appear to be superior options. Specialization among leaders also is not uncommon. A community may have one or more individuals serving as arbitrators or mediators, and/or as religious, hunt, or war leaders, and so on.

Warriors

Joint production of mutual defense against enemies evolves if outside threats are perceived, and such threats often mean an important part of an individual's belief system will be "a concept of them and us" (e.g., tribalism, patriotism). In fact, an external enemy can strengthen incentives for intragroup cooperation (Ridley 1996, p. 174). There is no centralized authority in a customary-law system, however, so individuals cannot be forced to become soldiers. They have to be persuaded to take on this role. This explains the incentives to propagate cultural beliefs (rules) about honor from bravery and skill in warfare in primitive and medieval communities. Members of customary-law communities facing outside threats (e.g., spouses or potential spouses of warriors, fathers and mothers of potential warriors, elders who can no longer fight) have strong incentives to encourage young males to be brave and to be skilled in combat (Benson and Siddiqui 2014). Thus, the successful warrior is honored, but he also receives many personal rewards. A warrior who backs down from a threat or who does not pursue retribution will be seen as a coward, and this generally will be the greatest fear instilled in boys and young men. The prestige associated with abilities in warfare also encourages entrepreneurial war leaders to organize attacks on enemies. After all, given the belief that some other group is made up of enemies, aggression can easily be rationalized – "the best defense is a good offense" – particularly when the expected gains exceed the expected costs.

The discussion of conflict suggests a question: why warfare rather than cooperation (negotiation, diplomacy) with outsiders. In fact, as explained below, cooperation between members of different communities is also widely observed in customary-law systems. Given the high transaction costs for multiple-community collective action, however, the benefits of

intergroup cooperation have to be high for stable cooperation to develop (e.g., high enough for warriors to refrain from aggressive acts even when very attractive opportunities to attack arise). One potentially large benefit from intergroup cooperation arises when two or more groups face the same relatively powerful enemy. In this context, however, it should be noted that terms like "alliance," which suggest established protocol, permanence, and formality, do not describe military relationships between such communities. These cooperative activities generally are expedient combinations in which distinct and autonomous individuals and groups work toward common but limited aims. Indeed, military coalitions generally are temporary spontaneous arrangements, and they can fall apart when the common threat loses power and/or one of the groups in the coalition gain power relative to others (Benson 2006).

Conflicts between customary-law communities need not be violent. When two different commercial communities have disputes or try to capture each other's markets, they generally do not go to war with one another. Their warriors may be lawyers who pursue objectives through adversarial adjudication processes that arise, for instance, or lobbyists who compete for political favors in an authoritarian legal system, as discussed below. Conflict is certainly not inevitable, however, if benefits of intercommunity cooperation are significant.

Polycentric Law

Customary rules and procedures often facilitate voluntary cooperative interactions such as trade between members of different customary-law communities (Benson 1988). Groups need not formally "merge" and accept a common set of rules, however, in order to achieve intercommunity cooperation on some dimensions. Individuals only have to expect each other to recognize specific rules pertaining to the types of intergroup interactions that evolve. Indeed, a "jurisdictional hierarchy" may arise wherein each group has its own rules and procedures for intra-community relationships, while a separate more narrowly focused set of behavioral and/or procedural rules applying for intercommunity relations (Pospisil 1971; Benson 1988). For instance, intra-community recognition of rules is likely to be largely based on reciprocities, trust, and reputation, along with ostracism threats (prudent morality), while intergroup recognition may require bonding, strong surety commitments, and/or threats of retribution (retributive morality). A jurisdictional hierarchy also does not create a higher order of law, as intergroup institutions typically have no role in any community's internal relationships. Thus, customary law is "polycentric," with multiple parallel "local" jurisdictions, as well as overlapping jurisdictions supporting intercommunity interactions. Many intragroup rules will be common across different groups, of course, as individuals in different groups discover similar ways to deal with an issue. Emulation also will occur where differences initially exist but individuals perceive superior arrangements among other groups (Benson 1988, 1989).

Custom Versus Authority

Pospisil (1971) distinguishes between "legal" arrangements that evolve from the top down through command and coercion, which he called "authoritarian law," and customary-law systems that evolve from the bottom up through voluntary interaction. Similarly, Hayek (1973, p. 82) distinguishes between "purpose-independent rules of conduct" that evolve from the bottom up and rules that are designed for a purpose and imposed by "rulers." Individuals may be members of different specialized customary-law communities (e.g., diamond merchants (Bernstein 1992) or trade associations (Benson 1995), a neighborhood (Ellickson 1991), and so-called informal or underground communities (de Soto 1989)) while simultaneously being subjected to authoritarian law. Both Hayek's and Pospisil's distinctions suggest that customary law also may conflict with authoritarian law. Indeed, as Hayek (1973, p. 82) stresses, "the growth of the purpose-independent

rules of conduct which can produce a spontaneous order will . . . often have taken place in conflict with the aims of the rulers who tended to turn their domain into an organization proper." When this occurs, an authority backed by coercive power may attempt to assert jurisdiction over a customary-law community, but this will have very different impacts, depending on the options available to members of the community. If the authority is strong enough, a community might be forced to accept the commands, although new customary rules and procedures often evolve that allow a customary-law community to avoid some and perhaps most authoritarian supervision (Bernstein 1992; Benson 1995). This may involve moving "underground," making the customary rules and procedures difficult to observe, and raising the cost of enforcement for the authority. If a customary-law community (and its wealth) is geographically mobile, however, members may simply move outside an authority's jurisdiction. The threat to move can significantly constrain authoritarian attempts to displace a customary system (Benson 1989). A sovereign who wants to avoid an exodus may even offer to assist in enforcing the customary rules, and even explicitly codify them. Indeed, a sovereign may offer special privileges to members of a highly mobile customary community in order to capture benefits including revenues, perhaps directly from tribute or taxes, but also indirectly because of a positive impact that this community has on other less-mobile sources of wealth (e.g., land) that can more easily be controlled and taxed. If customary behavioral rules are absorbed and customary procedures atrophy, an authority may amend or replace many customary rules, although the ability to do so depends on the costs of reinvigorating customary institutions, the mobility of wealth for members of the community, and the privileges granted to the community's members.

Cross-References

► Lex Mercatoria

References

Barton RF (1967) Procedure among the Ifugoa. In: Bohannan P (ed) Law and warfare. Natural History Press, Garden City, pp 161–181

Benson BL (1988) Legal evolution in primitive societies. J Inst Theor Econ 144:772–788

Benson BL (1989) The spontaneous evolution of commercial law. South Econ J 55:644–661

Benson BL (1991) An evolutionary contractarian view of primitive law: the institutions and incentives arising under customary American Indian law. Rev Austrian Econ 5:65–89

Benson BL (1994) Are public goods really common pools: considerations of the evolution of policing and highways in England. Econ Inquiry 32:249–271

Benson BL (1995) An exploration of the impact of modern arbitration statutes on the development of arbitration in the United States. J Law Econ Organ 11: 479–501

Benson BL (2006) Property rights and the buffalo economy of the Great Plain. In: Anderson T, Benson BL, Flannagan T (eds) Self determination: the other path for Native Americans. Stanford University Press, Palo Alto, pp 29–67

Benson BL (2011) The enterprise of law: justice without the state, 2nd edn. Independent Institute, Oakland

Benson BL (2014) Customary commercial law, credibility, contracting, and credit in the High Middle Ages. In Boettke P, Zywicki T (eds) Austrian law and economics. Edward Elgar, London (forthcoming)

Benson BL, Siddiqui ZR (2014) Pashtunwali – law of the lawless and defense of the stateless. Int Rev Law Econ 37:108–120

Bernstein L (1992) Opting out of the legal system: extralegal contractual relations in the diamond industry. J Leg Stud 21:115–158

de Soto H (1989) The other path: the invisible evolution in the third world. Perennial Library, New York

Ellickson RC (1991) Order without law: how neighbors settle disputes. Harvard University Press, Cambridge, MA

Friedman D (1979) Private creation and enforcement of law: a historical case. J Leg Stud 8:399–415

Fuller L (1981) The principles of social order. Duke University Press, Durham

Goldsmidt W (1951) Ethics and the structure of society: an ethnological contribution to the sociology of knowledge. Am Anthropol 53:506–524

Hayek FA (1973) Law, legislation, and liberty, volume I: Rules and order. University of Chicago Press, Chicago

Johnsen DB (1986) The formation and protection of property rights among the Southern Kwakiutl Indians. J Leg Stud 15:41–68

Kreps DM (1990) Game theory and economic modeling. Oxford University Press, Oxford, UK

North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge, UK
Parisi F (2001) The genesis of liability in ancient law. Am Law Econ Rev 3:82–124
Pospisil L (1971) Anthropology of law: A comparative theory. Harper and Row, New York.
Ridley M (1996) The origins of virtue: human instincts and the evolution of cooperation. Viking Penguin, New York
Vanberg VJ, Buchanan JM (1990) Rational choice and moral order. In: Nichols JH Jr, Wright C (eds) From political economy to economics and back. Institute for Contemporary Studies, San Francisco, pp 175–236
Vanberg VJ, Congleton RD (1992) Rationality, morality and exit. Am Political Sci Rev 86:418–431

Customary Law of Merchants

► Lex Mercatoria