T

Takings

Thomas J. Miceli and Kathleen Segerson
Department of Economics, University of Connecticut, Storrs, CT, USA

Abstract

This entry discusses the economics of eminent domain, which is the government's power to take or regulate privately owned property for the common good. It discusses the origins of the power as well as its limits, particularly as embodied in the public use and just compensation requirements. It also reviews the economics literature on how eminent domain affects incentives for efficient land use.

JEL codes: K11, K32

Definition

Takings: The acquisition of privately owned property by the government using its power of eminent domain.

The right of the community to exercise eminent domain and thereby expropriate individually held property for the common good seems to have been universally acknowledged and practiced by societies throughout history. As Reynolds (2010, p. 11) asserts, ". . .the principle that land might be taken from individuals when the community needed it has been so generally accepted that it did not need to be stated or argued about until recent times." For example, the very phrasing of the Takings Clause in the US Constitution, which states "nor shall private property be taken for public use, without just compensation," implicitly acknowledges that governments have the right to take property; its main purpose is to limit that power by requiring that the taking be for a public purpose and that just compensation be paid.

The idea that compensation should be paid when land is taken also has ancient roots and is therefore taken for granted in the law of most countries. Even in communist China, where land is collectively owned, the 1982 Constitution provides for farmers to receive compensation for "use rights" (as proxied by crop yield) when their land is taken in the "public interest." In cases of physical takings, the debate about compensation has always been about the *amount* of compensation that should be paid, not about whether it should be paid. However, the question about the specific meaning of "common good" or "public purpose" seems to be primarily an issue in American law and likely reflects the explicit inclusion of "public use" as a limitation on eminent domain in the US Constitution.

Takings questions also arise in contexts where land is not physically taken, but its use is restricted in some way, for example, through land use regulations. Such regulations are often termed

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

"regulatory takings." In addition to the provisions for compensation for acquisitions under eminent domain, laws governing compensation for land use regulations exist in most countries, though they tend to vary greatly in the rights to compensation that they afford. In a study of 13 countries, including the USA, Canada, several European countries, Israel, and Australia, Alterman (2010) found a range of compensation rights, from very little to nearly full, with the USA lying in between. Notably, there seems to be no way to correlate a country's laws with its geographic, legal, demographic, or other attributes.

From a theoretical perspective, the issue of eminent domain has generated an enormous amount of legal scholarship and case law. Much of the discussion focuses on the meaning of the "public use" and "just compensation" requirements in the US Constitution, which together restrict the scope of government acquisition of private property. However, the economic questions regarding the appropriate limits on the power to take property or restrict its use are relevant in all countries where governments are granted these powers. The economic literature asks whether there is an economic rationale for bestowing the right to acquire property or restrict its use without the owner's consent on the government alone (and not on private parties) and seeks to understand the limits of that right, as embodied in the public use and just compensation requirements. The primary focus is on how granting and restricting that right affects incentives for efficient land use.

Public Use

On its face, a public use requirement would seem to limit the use of eminent domain to provision of public goods like highways, airports, or parks. This interpretation is appealing both in terms of the plain meaning of the phrase "public use" and in view of the well-accepted role of the government in providing these types of goods. Specifically, because public goods have the characteristic of non-excludability, meaning consumers can enjoy the benefits of the goods without first having to pay for them (i.e., to free ride), ordinary markets will underprovide them. Economists have long argued, therefore, that the government should provide public goods so as to ensure that the efficient quantity will be supplied and then use its tax powers to coerce consumers to contribute to the cost. In this sense, government provision and financing of public goods amount to a kind of "forced purchase" by consumers. Merrill (1986) refers to the proposition that the government alone should have the power of eminent domain and should only use it to provide public goods, as the "ends approach" to defining public use because it concerns the use to which the taken land will be put (also see Miceli (2011), Chap. 2).

The logic of the ends approach, however, does not explain why the land (and possibly other inputs) needed to provide a public good must also be forcibly acquired from the owner(s). The economic justification for this form of coercion, which involves a "forced sale" from input owners to the government, is a different kind of market failure, referred to as the holdout problem. The holdout problem arises in the context of any development project requiring the assembly of multiple contiguous parcels of land. The difficulty developers face in this context is that once the scope of the project becomes public knowledge, individual landowners acquire a kind of monopoly power, given that each parcel is essential for completion of the overall project. Imagine, for example, a road or railroad builder who has decided on the optimal route and has assembled several parcels. Refusal of any additional owners to sell would greatly increase the cost of the project, and as a result, all owners can hold out for prices above their true valuations. The likely result is underprovision of projects involving assembly (Miceli and Segerson 2012). One solution to the holdout problem is to take away an owner's right to refuse a sale at the offered price. The power of eminent domain represents such a forced sale at a price set by the court. Merrill (1986) refers to this justification for eminent domain – i.e., as a response to the holdout problem – as the "means approach" to public use because it concerns the manner by which land for a government project is acquired.

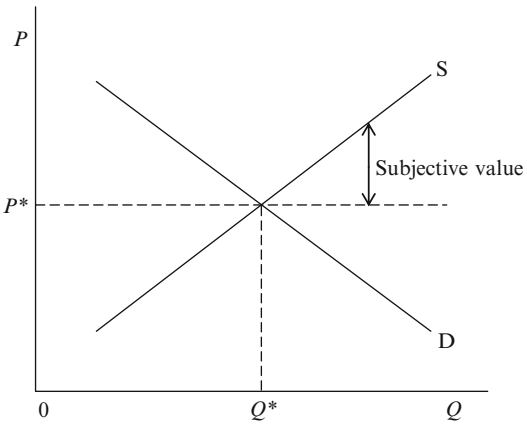
From an economic perspective, the ends approach is the correct justification for eminent domain, but that does not mean that it is the correct interpretation of the "public use" requirement. The reason for this paradox is twofold. The first derives from the fact that the holdout problem and the attendant inefficiencies are not unique to public projects. Private developers also face holdouts for large-scale commercial developments, as do local development authorities undertaking urban renewal. In some cases, private developers can mitigate the problem by means of secret purchases (Kelly 2006), but at some point the scope of the project becomes apparent, and the holdout problem arises. The logic of the above argument therefore implies that, from an efficiency perspective, the power of eminent domain should be available to any developer, private or public, engaged in land assembly.

Second, in explicating the constitutional basis for a public use requirement, courts routinely invoke the ends approach by appealing to the "public purpose" behind the project in question. This is an easy argument for truly public goods because the benefits of the resulting project are available to all, but the logic is less persuasive when the project is largely private, as has been true in most US Supreme Court cases involving the public use issue. For example, in the case of *Kelo v. New London*, despite the fact that the primary beneficiaries were private businesses, the US Supreme Court upheld the city's use of eminent domain to assemble land for purposes of a redevelopment project by emphasizing the jobs and enhanced tax revenues that would materialize from the project (125 S. Ct. 2655, 545 U.S. 469 (2005); also see Merrill (1986) and Kelly (2006) for a discussion of other notable cases in this vein). As Merrill (1986, p. 67) notes, however, this strategy is not unusual, as courts strain to make their reasoning consistent with the plain meaning of the US Constitution. The difficulty in defining the public use requirement therefore comes down to a divergence between the economic and legal justifications for eminent domain. Whereas economists focus on the use of eminent domain as an efficient response to the holdout problem, regardless of the context, courts often emphasize the need to identify some public benefits as flowing from the use of government coercion.

Just Compensation

A second requirement for the use of eminent domain is generally that the landowner must be paid "just compensation." Although no further definition of what amount of compensation is "just" is given, courts have typically interpreted it to mean "fair market value." Economists, however, argue that this measure systematically undercompensates owners because it ignores their "subjective value" (Fischel 1995a). The idea can be easily illustrated in a simple supply-and-demand diagram for some particular category of land (Miceli and Segerson 2007, p. 20) (Fig. 1).

In the graph, the demand curve represents the willingness to pay for individual parcels of land by potential buyers, while the supply curve represents the reservation price, or opportunity cost, of current owners. The equilibrium price, P^*, can be interpreted as the market value of a unit of the land in question, reflecting the price of recently sold parcels. Note that the price partitions the market into the parcels that are sold over some time period (those between 0 and Q^*) and those that are not sold (those to the right of Q^*). In other words, owners between 0 and Q^* voluntarily sold because the price exceeded their opportunity cost,



Takings, Fig. 1 Subjective value

but those to the right of Q^* did not sell because the price was below their opportunity cost.

Now suppose that one of the parcels to the right of Q^* is taken by the government for use in a public project, necessitating the payment of just compensation. In order for the transaction to be consensual on the part of the seller, compensation would have to be set at or above the relevant point on the supply curve. The problem with this "value-to-the-owner" measure of compensation is that it is a private information of the seller and therefore can be misrepresented, thus creating the risk of a holdout problem (Knetsch and Borcherding 1979). Market value, in contrast, is observable but clearly undercompensates owners, with the difference between market value and value to the owner representing the owner's subjective value, as shown in the graph.

Epstein (1985) has argued that the loss that market-value compensation imposes on landowners is only justifiable in two circumstances. The first is when the surplus created by the taking – the difference between the value of the parcel in its new use minus the owner's opportunity cost – is widely distributed rather than being concentrated in a few hands. In essence, Epstein is arguing for the ends approach to public use – that is, eminent domain should only be used to provide public goods – for in that case the distributional requirement is clearly satisfied (also see Ulen (1992)). Epstein's argument is motivated primarily by concerns about distributional equity, but it also reflects the idea that when gains from coerced transactions are dispersed, private interests will not find it worthwhile to engage in rent-seeking efforts to acquire eminent domain power.

The other circumstance in which undercompensation in a monetary sense is justifiable according to Epstein is when the government action provides "in-kind" compensation. An example is a zoning ordinance that deprives owners of some uses of their land, particularly those that would be harmful to their neighbors. Although an owner could claim that this type of regulation prevents him from engaging in certain profitable activities, like opening a gas station in a residential area, that "loss" is calculated based on his unilateral departure from an efficient land use pattern in which all other owners refrain from engaging in the externality-producing activity. In fact, when all landowners comply with the regulation, and assuming the regulation is efficient, their property values are enhanced relative to a situation in which no regulations are imposed. In this sense, all owners receive in-kind compensation from the regulation and therefore are not entitled to further monetary compensation. This is in contrast to regulations that "single out" individual owners to surrender their land for the benefit of others, in which case monetary compensation is required.

Compensation and Land Use Incentives

The preceding discussion of compensation has focused on its role in limiting excessive takings. However, most recent economic scholarship on eminent domain, especially since the publication of the seminal article by Blume et al. (1984) (hereafter BRS), has been to evaluate the incentive properties of the compensation rule regarding the land use decisions of property owners whose land is at risk of a taking. The incentive effects of compensation can be shown using a simplified version of the BRS model. Specifically, let $V(x)$ be the value of a piece of land after x dollars have been invested in improvements, where $V' > 0$ and $V'' < 0$. Also let p be the probability that the land will be taken by eminent domain, let B be the value of the land in public use, and let $C(x)$ be the amount of compensation paid to the landowner, which may depend on the amount of improvements. Initially, both p and B will be treated as exogenous.

The timing of events is as follows. First, the landowner chooses the level of investment, which is irreversible, taking the probability of a taking and the compensation rule as given. Once x is in place, the taking decision is made. If the land is taken, the landowner is paid compensation and the land is converted to public use. The requirement that x must be chosen before the taking decision occurs is crucial because otherwise, the landowner could simply wait until the taking decision is made and only invest if the land is not taken. (This assumption is not restrictive in the sense that private land is always subject to a taking risk.)

Similarly, the irreversibility of x precludes salvaging the cost of the investment if the land is taken.

In this setting, the socially optimal level of investment maximizes the expected value of the land:

$$(1-p)\,V(x) + pB - x. \tag{1}$$

The first-order condition defining the optimal investment, denoted x^*, is

$$(1-p)\,V'(x) - 1 = 0. \tag{2}$$

Note that the resulting investment is decreasing as the probability of a taking increases and is generally less than what the owner would invest in the absence of a taking risk, reflecting the fact that any improvements are lost if the land is converted to public use.

The actual level of investment by the owner maximizes the expected private value of the land, which includes the expected amount of compensation:

$$(1-p)\,V(x) + pC(x) - x. \tag{3}$$

The resulting first-order condition is

$$(1-p)\,V'(x) + pC'(x) - 1 = 0. \tag{4}$$

Comparing this to Eq. 2 shows that $C' \equiv 0$ is necessary and sufficient for efficiency – that is, compensation must be lump sum. Intuitively, compensation potentially creates a moral hazard problem. If, for example, landowners expected to be fully compensated, or $C(x) = V(x)$, they would overinvest because they would ignore the possibility that the land might be taken, which would render any improvements worthless. At the other extreme, zero compensation, or $C(x) \equiv 0$, results in the efficient level of investment. This represents the famous "zero compensation result" from BRS. (Of course, any positive lump-sum amount would also be efficient.)

The conclusion that zero compensation is efficient, however, is at odds both with general notions of fairness and constitutional requirements for just compensation. This raises the question of whether this result would hold under a more realistic depiction of the taking decision. For example, the above argument assumes that the probability that the land will be taken is independent of its private value. In reality, a government that behaves in a benevolent or "Pigovian" way (see Fischel and Shapiro 1989) will only take land when it is efficient to do so, thereby making the probability of a taking dependent on the landowner's decision. Suppose, for example, that the value of the land in public use is a random variable whose value is only observed after landowners have made their investment decisions. Upon observing the realized value of B, the government takes the land if $B \geq V(x)$, that is, if the land is more valuable in public use, given its current value to the owner. The resulting probability of a taking from the landowner's perspective, prior to the choice of x, is now $1 - F(V(x))$, where $F(V(x))$ is the probability that $B \leq V(x)$, with $F' > 0$.

The expected social value of the land in this case is

$$\begin{aligned} &F(V(x))\,V(x) \\ &\quad + [1 - F(V(x))]E[B|B \geq F(V(x))] \\ &\quad = F(V(x))\,V(x) + \int_{V(x)}^{\infty} BdF(B) - x, \end{aligned} \tag{5}$$

and the first-order condition defining x^* is

$$F(V(x))\,V'(x) - 1 = 0, \tag{6}$$

which is the analog to Eq. 2 with $F(V(x)) = 1 - p$. The expected private value is

$$F(V(x))\,V(x) + [1 - F(V(x))]C(x) - x, \tag{7}$$

and the first-order condition defining the privately optimal level of investment is

$$\begin{aligned} &F(V(x))\,V'(x) + [1 - F(V(x))]C'(x) \\ &\quad + F'(V(x))\,V'(x)[V(x) - C(x)] - 1 = 0. \end{aligned} \tag{8}$$

Note that $C' \equiv 0$ is no longer sufficient for efficiency, but it is still necessary. Now, in addition to being lump sum, compensation must also equal the value of the land at its efficient level of investment; that is, $C = V(x^*)$. The reason for this additional requirement is that landowners view the probability of a taking as depending on their

level of investment. In particular, by investing more (less), they can reduce (increase) that probability of a taking because it becomes less (more) likely that $B \geq V(x)$. Consequently, if compensation is less than full ($C < V(x)$), owners will overinvest to decrease the probability of a taking, and if compensation is more than full ($C > V(x)$), they will underinvest to increase that probability. Combining this with the lump-sum requirement yields $C = V(x^*)$ (Miceli 1991).

Hermalin (1995) shows that two other compensation rules will also achieve the efficient outcome in this context. The first requires the government to pay owners the full value of the land in public use, or $C = B$. In this case, landowners invest efficiently because they internalize the full social value of the land. Alternatively, compensation could be set at zero, but the landowner could be given the option to "buy back" the land in the event of a taking for a price equal to the realized value of the land in public use, B. The difference between these two rules depends on the assignment of property rights in the land when it potentially has public value. Under the first rule, the landowner holds the right, whereas under the second, society holds the right.

The preceding discussion assumed that the government behaved in a socially benevolent way in making its taking decision, but many have argued that an important reason for requiring compensation for takings is to prevent the government from converting too much land to public use (Johnson 1977). A government that considers the dollar costs of a taking as embodied in the compensation rule, rather than the true opportunity costs, is said to have "fiscal illusion" (BRS). To reflect this view of government behavior, suppose that once B is realized, a taking occurs if $B \geq C(x)$. It follows that "full" compensation will be necessary to prevent excessive takings, but simply setting $C(x) = V(x)$ will revive the moral hazard problem that initially gave rise to the BRS no-compensation result. One solution is the lump-sum rule, $C = V(x^*)$, which will simultaneously eliminate moral hazard by landowners (because compensation is lump sum) and induce only efficient takings by the government (because compensation is full).

Consider also the two rules proposed by Hermalin, both of which were shown above to induce efficient investment. Under the rule that sets $C = B$, the government will be indifferent between taking the land and not taking it for any realization of B. The landowner, however, will only want efficient takings to occur – that is, those for which $B \geq V(x)$. Thus, efficiency of the takings decision will depend on whether the government conforms to the wishes of the landowner. As for the buyback rule, since compensation is nominally set at zero, the government will initiate a taking for any $B > 0$, but the landowner will buy back the land whenever $V(x) \geq B$ or whenever a taking is not efficient. Thus, the outcome will be efficient.

A final compensation rule that balances the incentives of landowners and the government is the threshold rule proposed by Miceli and Segerson (1994). According to this rule, the government pays full compensation if it acts inefficiently to take (or regulate) the land, but it pays zero compensation if it acts efficiently. Formally, the rule is written as follows:

$$C = \begin{cases} V(x), & \text{if } B < V(x*) \\ 0, & \text{if } B \geq V(x*) \end{cases}. \tag{9}$$

The efficiency of this rule is established as follows. Assume the landowner invested efficiently. Then, once B is realized, if $B \geq V(x^*)$, the taking is efficient and compensation is zero. Thus, a government with fiscal illusion will incur a net benefit of $B > 0$ and so will go ahead with the taking. In contrast, if $B < V(x^*)$, a taking is not efficient and compensation is full. Thus the government will incur a net loss of $B - V(x^*) < 0$ and so will refrain from the taking. In both cases, the government acts efficiently. Moving back to the land use decision, the landowner, who anticipates the government's behavior, will choose x to maximize

$$F(V(x*))\,V(x) - x, \tag{10}$$

which has x^* as its solution. The Nash equilibrium therefore involves efficient behavior by both the landowner and the government.

The noteworthy feature of the compensation rule in Eq. 9 is that, in an efficient equilibrium, no compensation is paid for takings. This is clearly contrary to legal practice in takings cases involving physical acquisitions where compensation is typically required, but, as will be discussed below, it is consistent with not paying compensation in regulatory takings cases.

A final class of takings models, referred to as "constitutional choice" models, views landowners as designing the compensation rule from behind a veil of ignorance regarding which parcels will be taken for public use. In this setting, all landowners are equally at risk of having their land taken, given that it is efficient to devote some land to public use, but landowners also know that any money paid to takings victims must be raised by taxes assessed on all landowners. Thus, rational individuals will presumably account for both sides of the public budget in designing the compensation rule and will therefore not be overly stingy (in case their land is taken) or overly generous (in case it is not).

The prototypical model of this sort is by Fischel and Shapiro (1989) (also see Nosal (2001)), which is identical to the BRS model except that the probability of a taking is written as $p = s/n$, where n is the total number of parcels in the jurisdiction and s is the number that will be randomly taken. The public benefit is written as $B(s)$, which will be enjoyed by all landowners (including those whose land is taken), where $B' > 0$, $B'' < 0$. Compensation is written as a fraction of property value, or $C = \alpha V(x)$, where α is a nonnegative constant, and per-person tax liability is also assessed as a fraction of property value, or $T = tV(x)$, where t is the tax rate. The resulting expected wealth of a landowner is

$$\begin{aligned}&(1-p)\,V(x) + pC + B(s) - T - x\\ &\quad = (1-p+p\alpha-t)\,V(x) + B(s) - x. \qquad (11)\end{aligned}$$

The landowner chooses x to maximize this expression, taking p, s, α, and t as given. The resulting first-order condition is

$$(1-p+p\alpha-t)\,V'(x) - 1 = 0. \qquad (12)$$

As for the government, if it has fiscal illusion, it will choose s, the number of parcels to take, to maximize

$$nB(s) - sC. \qquad (13)$$

The resulting first-order condition is

$$nB'(s) - C = 0. \qquad (14)$$

Equations 12 and 14 jointly determine the optimal behavior of landowners and the government, given the compensation rule, as reflected by α, and the tax rate t. As noted, these parameters are determined by citizens from behind a veil of ignorance to maximize overall welfare, subject to a balanced budget condition. The latter is given by $sC = nT$ or

$$p\alpha V(x) = tV(x) \qquad (15)$$

for any x. It follows that $p\alpha = t$, which immediately implies that $x = x^*$ in Eq. 12. Thus, landowners invest efficiently regardless of the value of α. Although both the compensation rule and the proportional property tax are potentially distortionary with respect to the choice of x, in the current model the two distortions exactly offset through the budget constraint, resulting in an efficient level of investment for any amount of compensation (Miceli 2008). Finally, substituting for C in Eq. 14 implies $nB'(s) - \alpha V(x^*) = 0$. If $\alpha = 1$, this becomes

$$nB'(s) = V(x*), \qquad (16)$$

which says that parcels should be taken until the aggregate benefit from the last parcel taken just equals the opportunity cost. This, of course, is the Samuelson condition for efficient provision of a pure public good. Thus, if compensation is full, the efficient level of the public good is provided. This, therefore, is the choice citizens would make in designing the optimal compensation rule.

Regulatory Takings

Much more pervasive than physical acquisitions are government regulations that reduce the value of land without seizing title to it. Examples include zoning, environmental and safety regulations, and historic landmark designations. As noted, many courts have granted the government broad police powers to enact such regulations in the public interest without requiring them to compensate landowners, but in some cases the regulation goes so far in reducing the value of the regulated land that courts have declared it to be a regulatory taking for which compensation is due. The question is where the dividing line should be between compensable and non-compensable regulations (Michelman 1967; Fischel 1995b).

From an economic perspective, there is no substantive difference between a government action that seizes land for purposes of providing a public good and one that merely regulates that same property for purposes of preventing an external harm. In both cases, the government imposes costs on some landowners in order to confer benefits on others. The only difference is the extent of the taking. This point is clear in the context of the above models, which apply equally to full or partial takings. From a legal perspective, however, the two types of cases are often treated very differently by courts – as noted, full takings typically require payment of compensation, whereas partial takings do not. This dichotomy poses a significant challenge for developing a positive economic theory of the compensation question.

One answer is provided by the threshold compensation rule in Eq. 9, which, recall, only requires compensation to be paid for government actions that are inefficiently imposed. In this sense, the rule resembles the famous test for compensation announced by the US Supreme Court in the landmark case of *Pennsylvania Coal v. Mahon* (260 U.S. 393, 1922). According to this rule, referred to as the "diminution of value" test, compensation is only due if a regulation "goes too far" in reducing the landowner's value, where, in light of Eq. 9, "too far" can be interpreted as "inefficient." The efficiency of the Nash equilibrium under this rule, as demonstrated above, implies that compensation will be rarely paid, which is consistent with legal practice in many regulatory takings cases.

Another perspective on the distinction between full and partial takings as regards compensation is Epstein's idea, discussed above, that compensation need not always be monetary – it can also be in kind. Regulations that are enacted for the purpose of preventing externalities impose costs on individual landowners by limiting those things they can do with their property, but they also confer reciprocal benefits that serve as in-kind benefits. If the regulation is efficient, the benefits exceed the costs on average, which justifies nonpayment of monetary compensation. In contrast, physical takings are more likely to "single out" a small number of landowners to bear costs for the benefit of many, as when land is taken from a few to provide a public good. Since in-kind compensation is not generally available to these takings victims, monetary compensation is necessary to satisfy a just compensation requirement.

Acknowledgments We acknowledge the input of James Wen.

References

Alterman R (2010) Takings international: a comparative perspective on land use regulations and compensation rights. American Bar Association, Chicago

Blume L, Rubinfeld D, Shapiro P (1984) The taking of land: when should compensation be paid? Q J Econ 99:71–92

Epstein R (1985) Takings: private property and the power of eminent domain. Harvard University Press, Cambridge, MA

Fischel W (1995a) The offer/ask disparity and just compensation for takings: a constitutional choice approach. Int Rev Law Econ 15:187–203

Fischel W (1995b) Regulatory takings: law, economics, and politics. Harvard University Press, Cambridge, MA

Fischel W, Shapiro P (1989) A constitutional choice model of compensation for takings. Int Rev Law Econ 9:115–128

Hermalin B (1995) An economic analysis of takings. J Law Econ Organ 11:64–86

Johnson M (1977) Planning without prices: a discussion of land use regulation without compensation. In: Siegan B (ed) Planning without prices. Lexington Books, Lexington
Kelly D (2006) The 'public use' requirement in eminent domain law: a rationale based on secret purchases and private influence. Cornell Law Rev 92:1–65
Knetsch J, Borcherding T (1979) Expropriation of private property and the basis for compensation. Univ Toronto Law J 29:237–252
Merrill T (1986) The economics of public use. Cornell Law Rev 72:61–116
Miceli T (1991) Compensation for the taking of land under eminent domain. J Inst Theoretical Econ 147:354–363
Miceli T (2008) Public goods, taxes, and takings. Int Rev Law Econ 28:287–293
Miceli T (2011) The economic theory of eminent domain: private property, public use. Cambridge University Press, New York
Miceli T, Segerson K (1994) Regulatory takings: when should compensation be paid? J Leg Stud 23: 749–776
Miceli T, Segerson K (2012) Land assembly and the holdout problem under sequential bargaining. American Law and Econ Rev 14:372–390
Miceli T, Segerson K (2007) The economics of eminent domain: private property, public use, and just compensation, vol 3, Foundations and trends in microeconomics. Now Publishers, Boston
Michelman F (1967) Property, utility, and fairness: comments on the ethical foundations of 'just compensation' law. Harv Law Rev 80:1165–1258
Nosal E (2001) The taking of land: market value compensation should be paid. J Public Econ 82: 431–443
Reynolds S (2010) Before eminent domain. University of North Carolina Press, Chapel Hill
Ulen T (1992) The public use of private property: a dual constraint theory of efficient government takings. In: Mercuro N (ed) Taking property and just compensation: law and economics perspectives of the takings issue. Kluwer, Boston

Tariff Benefits

► Preferential Tariffs

Tariff Concessions

► Preferential Tariffs

Tax Amnesty

Carla Marchese
Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, POLIS Institute, Università degli Studi del Piemonte Orientale "Amedeo Avogadro", Alessandria, Italy

Definition

Tax amnesty is the opportunity given to taxpayers to write off an existing tax liability (including interests and fines) by paying a defined amount. Such offers are usually presented as being exceptional and available for only a limited period of time. Amnesties can either be general or restricted to certain groups of taxpayers or taxes, and they routinely include the waiving of criminal and civil penalties.

Prevalence and Types of Tax Amnesties

Both local and central authorities grant tax amnesties. Over the past 50 years, the central governments of some developing countries (e.g., Argentina, Colombia, Brazil, India, the Philippines, Turkey) have repeatedly offered amnesties (Le Borge and Baer 2008), as have the central governments of developed countries plagued by specific economic problems such as recession, financial crisis, and large public debt (e.g., Ireland, Italy, Spain, Greece, Portugal). Many developing and developed countries have also occasionally resorted to some form of amnesty to foster flight capital repatriation or to ease economic liberalization and openness to international trade. Local governments, too, often resort to tax amnesties. Many US states have made repeated use of waves of amnesties (Alm and Beck 1993; Mikesell et al. 2012) in response to a variety of motivations, including decreased central support for local tax enforcement in the 1980s (Dubin et al. 1982) or the dwindling of local tax revenue coupled with a mandatory balanced budget in the 2000s.

Lawmakers and local administrators are constantly devising new types of tax amnesties. Innovation and differentiation in this field is likely sparked by the need to capture the attention of the public and, where tax amnesties are frequent, to appeal to groups that have not yet been reached by previous offers.

In terms of the immediate financial benefits to participants, some tax amnesties not only reduce or waive sanctions and interest but also reduce the principal on the tax. These are the so-called extensive tax amnesties (Franzoni 1996; Macho-Stadler et al. 1999), which also often provide immunity from audits for past, and sometimes future, tax liabilities.

The timing of amnesties is a key feature in their functioning: how long the program is available, whether or not extensions will be granted, and the frequency with which amnesties are offered clearly affect their results (Mikesell 1986). In many countries, unaudited taxpayers who spontaneously report tax evasion can be granted a standing permanent tax amnesty (Andreoni 1991), although these amnesties are never of the extensive type and are sometimes available only for a limited time after the violation. Standard tax amnesties instead can be designed to cover recent or past liabilities that still fall within the expiration term of the tax obligation. The benefits of amnesties for the participants sometimes also extend to the future, as various provisions can be introduced to reduce expected future liabilities. For example, in some amnesties, participants who increased their subsequent reported income by a given percentage for several years were exempted from future audits on those years, barring major violations. Portugal granted an amnesty of this type in 1986 (Baer and Le Borgne 2008, p. 10).

Another important aspect of tax amnesties is the information disclosed by participants, which serves to condition their future expected payments. Some amnesties even provide for the free writing-off of past liabilities, so long as the taxpayer's latest tax return was honest (Pommerehne and Zweifel 1991). When a new tax return is filed, the tax administration ordinarily maintains its full powers of auditing. Taxpayers participating in an amnesty may also be subject to special surveillance in subsequent years. Of course, these policies reduce the amnesty's appeal and its potential as an immediate revenue source. If the government is primarily interested in raising revenue and encouraging participation to boost the immediate amnesty's proceeds, the auditing powers can be limited or excluded, and anonymity can be offered to the amnesty participants. One way this can be achieved is by allowing participants to disclose their liabilities and to make the amnesty payments to a third party (such as a bank), which releases a certificate to be used as a shield in case of future tax audits: the 2001 Italian tax amnesty for capital repatriation with these characteristics was called a "scudo fiscale" (tax shield). However, it is also true that the government's commitment not to access such information may be more or less credible. In Italy, a 2011 law introduced a new tax on capital that benefited from that 2001 amnesty. The new tax was justified on the basis of the benefit principle, with the benefit being continued secrecy to those who entered the amnesty despite newly introduced legislation granting the tax administration easier access to taxpayers' bank accounts.

In terms of the extent of coverage, amnesties are often granted only to taxpayers not yet under investigation. These could be taxpayers who missed a filing deadline (e.g., for VAT), or who failed to file one or more tax returns, or simply those who reported regularly but cheated. However, amnesties can also include those whose liability has already been assessed or liquidated (i.e., the so-called accounts receivable). By granting an amnesty, the tax administration surrenders the right to collect payments from taxpayers through standard means, such as audits, injunctions, or litigation in courts. The ensuing opportunity cost is likely to be larger if the ordinary collection process has already been initiated, as in the case of accounts receivable. Another aspect of coverage is the type of tax or tax base to which the amnesty refers. From this point of view, in principle, all types of payments can be considered, including social contributions, charges and fees, and so on. Amnesties tend to be general to assuage taxpayer fears that further audits will ensue if one specific hidden tax base is disclosed. With

reference to the tax base, it is also important to distinguish standard amnesties from those specifically targeted to flight capital, since the latter involve problems of international relations and tax competition. Amnesties can also involve leverage, so in some cases participants have been required to invest the hidden tax base in special public debt bonds. These securities delivered no yield or a low yield and could not be traded before a given date. As in the case of intermediation by third parties, these special bonds could be used as a shield in case of tax audits. An amnesty of this type was granted in Spain in 1991 (Macho-Stadler et al. 1999). In other cases (e.g., the 1987 amnesty in Argentina) the evaded tax base could also be invested in private enterprises, provided that the investment was twice the evaded tax base amount.

Amnesties are also characterized by the interventions supporting them, which range from global reforms of the tax system to specific provisions aimed at strengthening tax enforcement. The underlying reasoning is some stick should be given along with the carrot represented by the amnesty, in order to avoid negative effects on compliance. These further interventions often include harsher penalties for tax evasion, reorganization of the activities and of the legal capabilities of tax auditors, modifications of laws regulating tax shelters, and use of the funds collected through the amnesty for financing enforcement activities. Moreover, specific information can be sent to the perspective participants to inform them of their likely tax liability, together with the threat of naming and shaming evaders who do not participate or of increasing penalties specifically for them.

A traditional justification often provided for granting a tax amnesty is that special circumstances may motivate unwanted breaches of the law or mistakes by citizens. This is mainly true for amnesties that accompany huge reforms in taxation or other related fields or that are granted after major upheavals such as political regime change, changes in currency, and so on. In these cases, the amnesty is well grounded in terms of equity and should not be harmful in terms of efficiency, since it should be unexpected (and will not induce an ex ante evasion increase) and unlikely to be repeated (and will not encourage subsequent evasion). The resort to tax amnesties, however, is more frequent and widespread than one would expect on the basis of exceptional circumstances alone. The Philippine government, for example, called its 1980 amnesty the "final amnesty," although many others followed.

Amnesties are actually a well-established, ancient, and widely used institution. There is evidence of a tax amnesty in the Rosetta stone (from around 200 B.C.): the priests of a Memphis temple thank the Monarch for not demanding a large sum of tax arrears due by the people. Tax amnesties can also be considered a form of pardon that shares some features with other past or recent institutions (Cassone and Marchese 1999) in the field of religion (jubilees, indulgences), criminal justice (plea bargaining), and penal or civil violations (amnesties for illegal immigration, breach of regulatory rules, and so on). The provision of statutes of limitation for some legal obligations and even for crimes can also be considered as a limiting case of a permanent standing amnesty, which might be rationalized on the basis of the fact that as time passes without any actual law enforcement, the net advantages of delayed enforcement tend to vanish or even turn to disadvantages. These considerations suggest that amnesties might exert some positive social function. However, since tax amnesties are also harshly criticized, it is important to understand both the pros and the cons of this institution.

Pros

Among the pros is the idea that amnesties encourage repentance of violators and/or foster their willingness to behave cooperatively in the future (Malik and Schwab 1991). People may be unable to clearly assess ex ante the costs and benefits of violating rules, such as engaging in tax evasion, and while ex post repentance may ensue, the fear of heavy sanctions for past conduct sometimes discourages disclosure of the violation. Amnesties have the positive effect of rendering repentance less costly and the return to honest behavior (such

as the regular payment of future taxes) therefore more likely. The benefits of amnesties in this case are the partial restoration received by society (through payments or other forms of contribution by the violators) and by expected improvement in future compliance.

While repentance implies some form of limited rationality, the full rationality of taxpayers can also be assumed when justifying the participation of citizens and the granting of tax amnesties by governments. Participation can be justified if it is in the best interest of the taxpayer to revise her economic calculus because the costs of evasion have increased, as a result of actual or anticipated changes in law enforcement, for example, which increase the expected sanctions for past misconduct. These revisions, however, are more likely for small evaders that can easily adapt to even marginal changes in the law, while the effects on repeat evaders who have hidden large sums are probably limited.

Good economic news might also provide the motivation for granting an amnesty. If the economy is growing quickly thanks to policies of liberalization and of opening to international trade, it may be the case that firms can benefit more in the new environment if they are legal and have a clean tax record, as this paves the way for accessing the credit market or for listing on the stock exchange and so on. For a firm that has been operating for some time in the hidden economy, however, shifting to legality could imply a huge cost in terms of sanctions for past evasion, and an amnesty can grease the wheels of change. The results of the amnesty in this case should be evaluated not only with reference to the effects on tax revenue but also in terms of the effects on GDP that should ensue thanks to productivity increases in firms that were able to shift to legality (Bose and Jetter 2012).

Amnesties can improve efficiency when they are designed as an optimal discriminatory policy in the field of taxation (Marchese and Cassone 2000). Amnesties are in general discriminatory since they imply a more favorable treatment for those who evaded tax – and for those who can regularize their position at a discount – than for those who complied from the outset on the one hand or for those who were discovered and punished before the amnesty on the other. Amnesties can perform a role similar to that of price discrimination which can increase a firm's profits. For example, selling a good in different markets at different prices can boost a firm's revenue, and this might occur even if some arbitrage arises; in other words, for example, even if a few of those who were expected to buy at a higher price manage to buy at a lower one. Amnesties can be seen as a way of opening another market beyond that of compliance. If such an offer appeals to a large enough number of new "customers", government revenue can increase even if the number of regular tax compliers decreases somewhat. Price discrimination, however, can be optimally designed only if there are characteristics that distinguish markets or, in our case, that distinguish perspective evaders and compliers.

For a case in which an efficient discrimination can be applied, consider economic shocks that affect particularly some firms or individuals (e.g., sectorial crises, adverse life events, and so forth). Tax evasion can work as an extreme method of increasing disposable income in such circumstances. Those more likely to be harshly hit by a negative shock are also those more likely to resort to such extreme measures. In this case, an amnesty helps the unluckiest to improve their circumstances as well as easing their return to legal behavior. Since there might always be a share of the population experiencing these problems, a standing permanent tax amnesty can be justified both in terms of efficiency (supplying insurance) and of equity (helping those more in need) (Andreoni 1991).

Another possible efficiency-based rationale for resorting to amnesties as a discriminatory policy is related to the exploitation of differences in the visibility of tax evasion (Marchese and Cassone 2000). Citizens who are more confident about their ability to remain undetected are less likely to file their tax returns regularly, but they may still be willing to pay to eliminate the risk of evading the law and thus be interested in an amnesty. While it might be difficult to distinguish ex ante those who are more difficult to tax, it should become clear ex post, since the more visible should comply immediately, while the less visible

dare to wait. If a favorable amnesty is offered, the latter can enter it. The tax administration bears an opportunity cost since without the amnesty it might have enforced sanctions on the evaders, but if the payment asked for entering the amnesty equals the expected sanctions plus a risk premium, the result could be positive in terms of government revenue. Note that this approach can justify expected, repeated, and periodic tax amnesties, which should work like a form of sales (Cassone and Marchese 1995). Sales, too, aim at discriminating among customers according to their impatience. It is well known that efficient periodic sales can be compatible with a market in equilibrium, in which the share of customers buying in periods in which the full price is applied is stable and the total revenue is larger than if sales are not held. Similarly efficient periodic tax amnesties are compatible with the stability of the number of regular compliers. A characteristic of both efficient amnesties and sales is the greater exploitation of those with the lowest demand (those who buy during sales or those who participate in a tax amnesty), while a better deal should be offered to regular buyers or compliers. This might seem somewhat counterintuitive, but the idea is that those who buy during sales and those who participate in amnesties, while paying less in absolute terms, should be fully expropriated of their willingness to pay, while those characterized by high demand (regular buyers and compliers) should be left with some net gain.

When amnesties are offered to tax evaders already under investigation, they can be rationalized as a form of plea bargaining. In other words, the tax administration uses the amnesty to profitably renounce part of the expected proceeds as long as this also involves a partial cashing of its credits and a substantial reduction in the implementation costs it would have borne. As in the case of plea bargaining, it is expected that those who are more willing to participate in the amnesty are also those who more patently violated the law, since they would lose if they tried to contest their liability. It is widely held that plea bargaining can be considered as an efficient selection tool, as on the one hand the guilty, who risk more, should reveal themselves by pleading guilty and thus be properly sanctioned, while by not pleading guilty, the innocent should go on to trial, where they are likely to be acquainted (Grossman and Katz 1983). A similar type selection is possible through amnesties. However, some problems may ensue with tax evasion, as long as only monetary sanctions are foreseen. In this case, *ceteris paribus*, the amnesty is more appealing to those who have enough wealth to bear the liability and are thus easier to reach through standard means of enforcement. An amnesty, instead, is not relevant for those who are sanction-proof, as they have nothing to lose if an audit occurs.

While efficient discrimination is a potential, overall justification for tax amnesties in both developed and developing countries, in the latter countries, a further motivation it that they can help compensate for organizational problems related to performing audits and dealing with taxpayers' tax-relate appeals. Here, the main idea is that, whenever the productive capacity of the tax administration is modest and fixed in the short run, performance can be improved by using the amnesty to deal with the oldest unsolved cases, in order to concentrate resources on the most recent and probably more visible and thus easier to resolve ones. Moreover, amnesties can also be used by governments to periodically curtail the rents extracted from evaders by corrupted auditors seeking bribes. As long as the amnesty provides a large enough discount, taxpayers will prefer to settle their liabilities directly with the state.

Cons

On the efficiency grounds, the main argument against tax amnesties is that credibility problems may arise concerning the ability and/or commitment of the state to enforcing the tax law (Stella 1991). The idea is that governments in general, when granting an amnesty, try to look tough for the future, in order to induce compliance. Citizens, however, base their beliefs on past experience. Even if amnesties are accompanied by declarations about new stronger enforcement efforts, taxpayers, who take past auditing policy into account, are likely to only slightly correct

their perception of the riskiness of tax evasion. Hence, the collection of resources from past evaders is likely to be low, and the expectation of future amnesties can induce previous compliers to evade. If an amnesty leads to larger evasion, a further subsequent amnesty might appear even more justified, but this might lead to a slippery slope, in which more and more generous and frequent amnesties are granted. Amnesties, in sum, could reveal the weakness behind the feigned tough stance of a government, thus permanently endangering its taxing capacity, i.e., reducing the overall taxpayers' willingness to pay, either through regular compliance or in amnesties.

As long as a government lacks the commitment to enforce the tax law, the aforementioned discriminatory function of amnesties would actually be impossible to implement due to the lack of credibility of the threat of punishment. Discriminatory policies have also been criticized because the tax administration might lack enough information to design them, if taxpayers' attitudes differ along many dimensions (such as income, age, sex, location, and so on). It might even be technically impossible to devise a sufficiently fine-grained discriminatory mechanism.

A further problem raised by amnesties is that they are costly, and their costs are difficult to anticipate and to correctly evaluate (Baer and Le Borgne 2008). As already seen, a large share of these costs is represented by the opportunity costs of renouncing pursuit of the standard enforcement activity over the participants, thus renouncing the revenue that they would otherwise have produced. When the amnesty foresees payment by installment, some installments may be missed, and the initial assessment of the amnesty's proceeds proves incorrect. Other costs can be due to a temporary suspension of enforcement activities, which is often granted during the period in which the amnesty is pending. Moreover, amnesties need to be well designed and advertised: "Get to us before we get to you" is the famous slogan used to advertise the Michigan 2002 tax amnesty. Interventions can include mass mailings of information to prospective participants, the provision of an information hotline, and so forth. Last but not least, if amnesties endanger the future compliance of honest taxpayers, this prospective revenue loss also needs to be accounted for. Difficulty in evaluating the total costs involved in an amnesty may actually contribute to their popularity, as they would thus be based on a kind of fiscal illusion.

In terms of equity, most critics are of the view that amnesties are unacceptable because they introduce discriminatory treatment of citizens according to law enforcement. More specifically, when an amnesty is granted *ceteris paribus*, the participants can fulfill their tax obligations by paying different amounts than those who complied from the outset or from those who were caught in the meantime. On the other hand, if one focuses not on the equity of rules but on the social outcome instead, one sees that even without amnesties, honest taxpayers and tax evaders routinely end up with differences in their actual contributions to the financing of the public budget. Moreover, as long as the amnesty collects revenue from tax evaders, these differences are reduced, thus securing more horizontal equity. Even vertical equity could increase, if the rich evade more and take advantage of tax amnesties more often. Here too, as is generally acknowledged, focusing on justice in procedures or on justice in outcomes leads to different conclusions.

Even by considering amnesties as discriminatory in principle, the economic approach would suggest considering the potential equity efficiency trade-off resulting from possible compensation for those who are negatively hit thanks to the proceeds that an efficient amnesty should produce. However, the compensation principle is problematic and a source of widespread debate: it would either have to involve the actual payment of compensations (and this would be difficult to plan for and is therefore practically never done) or a comparison on paper of utility gains and losses, which is not acceptable to those who claim there is no objective unit of measurement for making these calculations. At any rate, some types of amnesties, such as those aimed at capital repatriation or at easing the opening of the economy to international trade, have clear-cut and widely recognized general economic benefits extending beyond the effects on tax revenue. They are

therefore often considered more acceptable from an equity point of view.

Another criticism of amnesties is that external incentives might endanger internal incentives to legal behavior. In other words, as long as tax compliance in modern societies largely rests on the internal incentive represented by a moral imperative, the introduction of external incentives based on the gains that an amnesty can produce for both those who participate and those who do not could push the public reasoning away from the moral perspective. This might have negative effects on compliance, since in many cases, pure economic calculus shows that tax evasion pays, so widespread evasion should be expected. Amnesties, however, are generally considered as acceptable on the grounds of equity if they discriminate in favor of more deserving people: amnesties that work as a form of insurance for the most unfortunate or exceptional amnesties granted to citizens faced with the difficulties posed by large-scale overhauls of the tax code or other major changes.

Not everyone who files for an amnesty is a tax evader. Even honest taxpayers may fear having made mistakes on their tax returns and want to avoid being audited or going to court or they may also wish to avoid the inconvenience and cost of being audited. Particularly in developing countries, where enforcement is often intrusive and burdensome for citizens, amnesties can appeal to honest taxpayers. The role of amnesties in this case is ambiguous. On the one hand, they produce a benefit for participants who are honest and deserving from a social point of view; but, on the other, they can imply perverse incentives for the tax administration (Franzoni 2000) to capitalize on its own malfunctioning.

The Observed Effects of Amnesties

As for participation, amnesties tend to deliver somewhat extreme results. Either the entire potential population participates or only a very few members do. This is due to the fact that amnesties ordinarily reduce the perspective workload of auditors, who no longer need to focus on those who entered the amnesty and can thus target nonparticipants more. Hence, the risks for nonparticipants (and thus their motivation for taking part) increase with the number of those who have already applied. In a certain sense, there is a network effect, and even in successful amnesties, participation tends to accelerate toward the end of the period of validity, as people wait and see how large participation has been. It is thus suggested that the period in which an amnesty is pending should not be too long (around 90 days), in order to avoid wasting time and to reduce costs. Moreover, this period should not coincide with that in which tax reports are filled-in, since it has been shown that this might exacerbate the trade-off that might arise between reporting for the amnesty and reporting for paying taxes of the current year (Alm and Beck 1990). If both reports have to be made at the same time, when confronted with two ways for reducing their risk in the field of taxation, past evaders can more easily assess the relative advantages. If the amnesty is very cheap, they might even reduce the overall amount they pay, with a negative result for the public budget.

The economic effects of amnesties can be estimated using econometric techniques. The reliability of the results is limited by the fact that one either studies a specific amnesty – but then the data collection should be very detailed, and the results might at any rate not have general significance – or one can pool many amnesties, but then one needs to control for the specificities of each case considered. Moreover, it is not easy to disentangle the specific effects of amnesties from those of the supporting programs often jointly enacted.

Many studies have focused on the large number of amnesties granted in the USA, which present the advantage of providing a large data set of cases sharing a common legal and economic background. Mikesell and Ross 2012 arrive at a total of 117 amnesties granted by 41 States in the period 1980–2011. The results of these econometric studies show that some features of amnesties are decisive for boosting the proceeds, such as the effort made in advertising, the inclusion of accounts receivable, and the possibility for installment payments. Other relevant variables are those

linked to tax evasion opportunities, either in terms of self-employment or with reference to less auditing in the state by the central government. While the amnesties granted in the 1980s also included cases where eligibility was very restricted, the most recent amnesties are more often characterized by large admission criteria. Moreover, recent amnesties tend to waive interest to a larger extent (whereas the principal is never reduced) and are less often accompanied by supportive measures aimed at reinforcing future compliance. This might also be due to the repetition of amnesties over time, a fact that tends to reduce the number of new enforcement interventions not yet enacted that can be introduced and to undermine the credibility of further threats, so that they are mainly dispensed with. From the point of view of gross revenue collected, US states tax amnesties were sometimes successful (with an impact at any rate always below 3% of the yearly tax revenue), while their long-term effect is unclear and probably nil or negative (Alm and Beck 1993; Luitel and Sobel 2007). It has been noted that the participants totally unknown to the tax administration mainly continued to report after they entered an amnesty, but their contribution to the tax revenue has often been scant (Christian et al. 2002), thus raising some questions about the role of amnesties in significantly enlarging the future tax base. The average sums paid in amnesties were often small and related to recent years. Since there is no reason to expect this on an objective basis, the explanation may either be a lack-of-recall of past evasion or a rational calculus about the fact that evasion made far in the past and not yet discovered is even less likely to be found out in the future. In general, there seems to have been some change over time in the goals that the states pursued by granting an amnesty, which mainly ranged from boosting future compliance to providing an immediate source of revenue.

For countries that often granted tax amnesties (including Italy, Ireland, India, the Philippines, and Turkey), the best results seem to have been reached whenever an amnesty was offered together with a policy aimed at strengthening enforcement capacity and at improving the overall efficiency of the tax collection system (Baer and Le Borgne 2008). In some cases, instead, such as in Argentina, where measures of this type were not introduced, amnesties gave rise to a spiral in which forgiveness for not paying what was due according to a previous tax amnesty was also offered and participation in amnesties faded out over time. When a negative spiral is avoided, amnesties mainly modify somewhat the intertemporal profile of tax revenue; that is, they imply a temporary increase, also due to the fact that they anticipate the cashing in of some future revenue that would have been obtained through ordinary enforcement activity, followed by a subsequent decrease, with no change in the revenue trend.

Experiments have also been used to assess the effects of tax amnesties (Alm et al. 1990). It turns out that amnesties actually do foster subsequent increases in evasion. However, combining amnesties for the past with harsher penalties for the future prompts greater compliance. This combination is reminiscent of a feature that has also been deemed necessary for the efficient and equitable functioning of plea bargaining, where it is found that the judge should use discretionary power to threaten harsher penalties, so that those who plead guilty, while obtaining a discount compared to the new level of penalty, are actually given the sanction due (Grossman and Katz 1983).

Amnesties as a Public Choice Issue

Amnesties imply a kind of temporary modification of the social contract between citizens and the state on taxation. As with contract renegotiations, such modifications can be justified by the arrival of new information not available when the pact was signed or by changes in the preferences and objectives of the parties involved. It might, however, also arise out of intertemporal inconsistency, whenever the parties find it in their interest to renege on their past promises. In the specific case of taxation, the former case would occur with efficient amnesties that ease the adjustment to major changes or perform discriminatory tasks, while the latter would correspond to cases in which governments risk their credibility as law

enforcers in order to secure an immediate increase in revenue. The short-run perspective that often plagues the functioning of democratic governments can easily lead to intertemporal inconsistency. Notwithstanding the long-term benefits of tough tax enforcement, politicians who stay in power for a limited period are likely to be tempted by amnesties that grant immediately available proceeds, while they might not be so worried about the damage that will materialize only in the long run, after they step down. To avoid these drawbacks, in some countries, amnesties can only be introduced with the approval of large parliamentary majorities (as in Italy for amnesties waiving criminal offenses) or they are subjected to approval by referendum (as in Switzerland).

Amnesties can be tempting from the political perspective, too, because they represent means of gaining a quick increase in revenue without burdening the entire tax-paying population with changes in the tax law and rate increases. This might be particularly welcome in preelection periods, as politicians are keen on spending to boost economic growth and consent (the so-called political business cycle). Amnesties, however, might also be seen as a way of squeezing taxpayers, besides offending regular compliers. In fact, from the point of view of political consent, amnesties have not performed well. The US state governors granting one during the electoral year proved more likely not to be reelected (Le Borgne 2006).

Amnesties are relevant for politicians also when they participate in one. If information leaks and reaches the press, it can damage the social reputation and the possible political career of participants. Companies also risk damaging their reputation if it is discovered that they took advantage of an amnesty.

Economic psychology has characterized the implicit psychological contract between the state and the citizens concerning taxation. While citizens are obliged to pay taxes, the state must treat them respectfully yet punish those who fail to comply. If it does not punish them, the citizens who did comply may feel betrayed. In fact, in experiments where there are collective gains from cooperation, the participants often demand punishment for violators. In many cases, it turns out that those who cooperated are ready to sacrifice a share of their gains in exchange for implementing the punishment. If amnesties are perceived as a breach of the psychological contract of taxation, they will encourage further tax evasion. However, it is also true that honest taxpayers might consider participants in an amnesty as willing to change their behavior. Frustration that some evaders may go unpunished in amnesties can also be dealt with if the hidden evaders are threatened with harsher penalties. In fact, punishment can serve two main purposes: retribution for illicit conduct and restoration of legal order. While the retribution recouped via amnesties is lower than that provided for by standard rules, amnesties can convey some advantages in terms of restoration as long as they foster greater future compliance. From this point of view, amnesties should be favored by those who are more generally in favor of alternative penalties aimed at facilitating the social rehabilitation of those who breached the law (Rechberger et al. 2010). The ambiguous role that amnesties can play implies that public debate over an amnesty program may have important consequences. This is confirmed by experiments in Switzerland and in Costa Rica conducted by Torgler and Schaltegger in 2005. They found that only amnesties approved by referendum lead to increased compliance. This effect can be traced back to the formation of public opinion through the public discussions among participants in the experiment that accompanied the referendum. Participants perceived the amnesty not as an imposition from above but as an agreed-upon intervention with useful functions, and this in turn increased the social pressure for cooperation.

As for public opinion on amnesties, the Bank of Italy (Cannari and D'Alessio 2007) conducted interviews in 1992 and 2004 in which questions were asked about the government motivations for granting an amnesty, the results expected, and the respondent's evaluation of such a policy. Regarding the first question, the majority of respondents think that the Italian state resorts to amnesties either because it is powerless to punish evaders or because groups of evaders had lobbied for preferential treatment. Yet in reference to the evaluation of the consequences and the moral

judgment of this policy, only about 30% of respondents have clear-cut negative feelings (that evasion will increase and the policy offends honest citizens), while the remaining respondents express more nuanced opinions. The negative feelings, however, were more frequent in 2004 than in 1992, possibly due to some deterioration of the government's credibility in light of repeated tax amnesties.

Besides being relevant for internal political affairs, amnesties can also be linked to the state of cooperation or competition between governments or levels of government, since they represent a means of dealing with externalities in taxation and in enforcement policies. It is also the case that forms of imitation or competition often arise among neighboring countries, so amnesties sometimes spread from one country to another. The importance of externalities in this field is confirmed by an empirical analysis of the motivations leading states in the USA to grant an amnesty, which revealed that the likelihood of amnesties increased as the effort of the federal government in auditing taxpayers within the state decreased (Dubin et al. 1982; Le Borgne 2006). In the field of international relations, in 2010, the OECD suggested offshore voluntary disclosure programs as a solution to help governments benefit quickly in terms of revenue from the effects of improvements in international cooperation for information exchange and transparency that have occurred since the onset of the financial crisis. Voluntary disclosure implies a "limited-time offer by the government to a specified group of taxpayers to settle undisclosed or unpaid tax liabilities for a previous period in return for defined concessions over civil or criminal penalties. In some cases, there are also concessions over the amount of tax and/or interest payable" (OECD 2010, p. 11): the definition is very close to that of an amnesty.

International organizations such as the IMF have studied more in general the policy of tax amnesties (Baer and Le Borgne 2008). They arrive at a substantially negative evaluation of this institution. Their suggestion in this field is to avoid the resort to amnesties, while pursuing alternative policies instead, such as: (i) trying to reduce tax evasion by addressing its basic determinants (unsustainable tax system, insufficient and improper enforcement, the malfunctioning of courts, etc.); (ii) resorting to permanent programs for encouraging disclosure of tax evasion and for granting extended payment agreements to taxpayers under economic stress for personal or conjunctural reasons; and (iii) improving the functioning of the tax administration, for example, by granting it with the power of disposing of cases unlikely to lead to net contributions to the revenue. The basic idea is to consolidate taxpayers' expectations about the commitment of the state to fighting tax evasion, while also dealing with the problems that often motivate the granting of an amnesty in other ways. These policy suggestions have sound economic foundations. They can be likened to the commercial practice that uses price discrimination systematically rather than intermittently, so that, for example, special sales can be replaced by permanent offers at outlets specializing in major discounts. Following these suggestions, however, is not easy, particularly when one considers developing countries and countries where corruption is frequent. Whenever interventions such as a standing amnesty are introduced, it is possible that corrupt auditors will accept bribes in order to say that a taxpayer voluntary disclosed her evasion. Problems of this type have arisen in the past even in the USA (Andreoni 1991). Likewise, whenever a personalized deal (such as an individual installment plan for tax payments) must be designed, the risk of corruption of officials tends to be greater than when general public interventions such as tax amnesties are implemented, given that they are often regulated by the law. These considerations, coupled with the existence of genuine unanticipated phenomena that cannot be dealt with efficiently through other means or with discrimination opportunities not available elsewhere, suggest that amnesties are and are likely to remain an accepted tool in tax administration.

References

Alm J, Beck W (1991) Wiping the slate clean: individual response to state tax amnesties. South Econ J 57: 1043–1053

Alm J, Beck W (1993) Tax amnesties and compliance in the long run: a time series analysis. Natl Tax J 46:53–60

Andreoni J (1991) The desirability of a permanent tax amnesty. J Public Econ 45:143–159
Baer K, Le Borgne EL (2008) Tax amnesties. IMF, Washington, DC
Bose P, Jetter M (2012) Liberalization and tax amnesty in a developing economy. Econ Model 29:761–765
Cannari L, D'Alessio G (2007) Le opinioni degli Italiani sull'evasione fiscale, vol 618, Temi di discussione. Bank of Italy, Roma
Cassone A, Marchese C (1995) Tax amnesties as special sales offers: the Italian experience. Public Finance/ Finance Publiques 50:51–66
Cassone A, Marchese C (1999) The economics of religious indulgences. J Inst Theor Econ 155:429–442
Christian CW, Gupta S, Young JC (2002) Evidence on subsequent filing from the state of Michigan's income tax amnesty. Natl Tax J 55:703–721
Dubin JA, Graetz MJ, Wilde LL (1992) State income tax amnesties: causes. Q J Econ 107(3):1057–1070
Franzoni LA (1996) Punishment and grace: on the economics of tax amnesties. Public Finance 51: 353–368
Franzoni LA (2000) Amnesties, settlements and optimal tax enforcement. Economica 67:153–176
Grossman GM, Katz ML (1983) Plea bargaining and social welfare. Am Econ Rev 73:749–757
Le Borgne E (2006) Economic and political determinants of tax amnesties in the U.S. States. IMF WP 706/222
Luitel HS, Sobel RS (2007) The revenue impact of repeated tax amnesties. Public Budg Finance 27: 19–38
Macho-Stadler I, Olivella P, Pérez-Castrillo D (1999) Tax amnesties in a dynamic model of tax evasion. J Public Econ Theory 1:439–463
Malik A, Schwab RM (1991) The economics of tax amnesties. J Public Econ 46:29–49
Marchese C, Cassone A (2000) Tax amnesty as price-discriminating behavior by a monopolistic government. Eur J Law Econ 9:21–32
Mikesell JL (1986) Amnesties for state tax evaders: the nature of and response to recent programs. Natl Tax J 39:507–525
Mikesell JL, Ross JM (2012) Fast money? The contribution of state tax amnesties to public revenue systems. Natl Tax J 65:529–562
OECD (2010) Offshore voluntary disclosure, comparative analysis, guidance and policy advice. OECD, Paris
Pommerehne WW, Zweifel P (1991) Success of a tax amnesty: at the polls, for the fisc? Public Choice 72: 131–165
Rechberger S, Hartner M, Kirchler E, Hämmerle FK (2010) Tax amnesties, justice perceptions, and filing behavior: a simulation study. Law Policy 32: 214–225
Stella P (1991) An economic analysis of tax amnesties. J Public Econ 46:383–400
Torgler B, Schaltegger CA (2005) Tax amnesties and political participation. Public Finance Rev 33: 403–431

Tax Evasion by Firms

Laszlo Goerke
IAAEU (Institute for Labour Law and Industrial Relations in the European Union),
University Trier, Trier, Germany
IZA, Bonn, Germany
CESifo, Munich, Germany

Abstract

A standard finding in the analysis of tax evasion and avoidance by firms is that the decision about the firm's activity level can be separated from the evasion choice and vice versa, irrespective of the tax under consideration. The implications, relevant empirical evidence, and the robustness of this separability feature are surveyed. The article finishes with speculations about topics of future research with regard to tax evasion or avoidance by businesses.

JEL-Classification: H 25, H 26, K 34

Definition

Tax evasion (avoidance) by a firm represents the attempt to illegally (legally) reduce the payment of taxes which have to be remitted by a profit-maximizing entity to below the level prescribed by law.

Introduction

A substantial fraction of tax revenues in OECD countries is remitted by firms (OECD 2013). This is the case either because firms are legally obliged to pay, for example, corporate income, payroll, property, or consumption taxes, or because firms act as withholding agents, inter alia with respect to personal income taxes and social security contributions. Therefore, firms and businesses have ample scope for tax avoidance and tax evasion

activities. Slemrod (2007, p. 28), for example, reports that the average tax gap in the United States – that is, the difference between the amount of taxes due and the amount paid voluntarily and in time – was about 16% in 2001. Moreover, the tax gap for income taxes of small corporations and for business income was substantially higher. As a further piece of evidence, the VAT compliance gap is estimated to exceed 10% in a number of European Union member states (Keen 2013). Consequently, tax evasion and avoidance by firms is not only feasible but also seems to be widespread.

Despite this evidence, the vast majority of contributions on tax evasion has focused on and analyzed the decision by individuals to evade income taxes. (The relevant literature is surveyed, for example, by Andreoni et al. (1998), Alm (1999), Slemrod and Yitzhaki (2002), Marchese (2004), Franzoni (2009), and Sandmo (2012). The contributions by Cowell (2004) and Slemrod (2007) also include substantial sections on firms. See also ▶ "Tax Evasion by Individuals", this encyclopedia.) The extant literature on tax evasion by firms has then focused on results which are specific to these entities. Such features which differentiate a firm from an individual are, inter alia, (1) all payoffs can be defined in terms of money; (2) firms often have market power; (3) risk neutrality may be an appropriate approximation of preferences; and (4) within a firm, conflicts of interest between owners and managers are likely to arise.

Basic Model

Subsequently, we take up some of these features and analyze a firm's tax evasion and avoidance behavior. We consider a single firm which maximizes expected profits. In order to do so, the firm determines the level, x, of economic activity and the extent of tax evasion or avoidance. The overall rate of all taxes the firm has to remit is labeled t. If the firm does not evade or avoid taxes, its (legitimate) profits are given by $\pi(x, t)$. We assume that a profit-maximizing level of activity, x*, exists. This activity choice, x*, is defined by $\pi_x(x^*, t) = 0$ and $\pi_{xx}(x^*, t) < 0$, where subscripts denote partial derivatives. Taxes remitted by firms will generally reduce profits, implying that $\pi_t < 0$ holds. This will be the case unless firms (1) can fully shift taxes forward to customers or backward to employees or suppliers of input goods or (2) act as withholding agents. In such a situation, $\pi_t = 0$ will apply. Furthermore, a pure profit tax will generally not affect a firm's optimal activity choice ($\pi_{xt} = 0$). In general, however, higher taxes will reduce the incentives to exert an economic activity. This implies that the gain from higher activity shrinks with the tax rate ($\pi_{xt} < 0$).

If the firm avoids or evades taxes, it decides about the under-declaration of taxes, which is labeled E, such that the resulting monetary gain amounts to Et. We subsequently assume that taxes are under-declared ($E > 0$) and that E is less than the tax base, in contrast, for example to Virmani (1989) and Cremer and Gahvari (1992), who analyze models of tax evasion by firms which allow for corner solutions, i.e., outcomes in which either no tax evasion takes place or no taxes are paid. To ensure an interior solution, evasion or avoidance has to be costly. These costs are given by C(E, t, F), where C is increasing in the under-declaration, E, for $E > 0$ at an increasing rate ($C_E(0) = 0 < C_E(E > 0)$, $C_{EE} > 0$) and also rising in the tax rate, t ($C_t > 0$), and a parameter F, which can capture the penalty which a tax-evading firm has to pay if evasion is detected ($C_F > 0$).

In the case of tax avoidance, profits may be considered as certain with regard to the outcome of tax payments. If tax evasion takes place, there is a probability, p, that evasion is successful and profits amount to $\pi(x, t) + Et$ and a converse probability $1 - p$ that the firm is audited and evasion is detected, so that profits are given by $\pi(x, t) + Et - C(E, t, F)$. If the firm evades taxes, expected nonlegal profits hence equal

$$\pi^N(x, E) = \pi(x, t) + Et - (1 - p)C(E, t, F). \quad (1)$$

The maximization of nonlegal profits π^N implies two first-order conditions:

$$\pi_x = 0 \tag{2a}$$

and

$$t - (1 - p)C_E = 0. \tag{2b}$$

The optimal activity choice, x*, maximizes legal profits, π, while the optimal under-declaration, E*, balances the marginal gain in terms of lower tax payments with the greater expected costs of evasion or avoidance.

Major Findings

From conditions (2a) and (2b), a number of fundamental insights can be derived:

1. Analytically, there is no difference between tax evasion ($0 < p < 1$) and tax avoidance ($p = 0$). This is the case because a penalty or, alternatively, the costs of avoidance affect the firm's payoff, namely, expected profits, qualitatively in the same way. Consequently, from now on, we will only refer to tax evasion, although the exposition obviously applies to avoidance activities as well.
2. The firm's activity decision is separable from its evasion choice. That is, the model predicts, first, that the firm will always choose that level of economic activity, x*, which it would have selected in the absence of evasion activities. Second, the optimal extent of tax evasion, E*, is the same, irrespective of activity choices. Accordingly, large firms, which are characterized by a high activity level, under-declare the same amount as otherwise identical small firms which exhibit a lower activity level. Note that this separability feature will also arise if the audit probability, $1 - p$, depends on the under-declaration, such that $p = p(E)$, because the firm's optimal activity choice is still determined by (2a).

   The separability prediction is due to two features. First, the monetary gains from legal profits, π, and from tax evasion, $Et - (1 - p)C$, affect the firm's payoff in qualitatively the same way and do not reinforce or weaken each other. This is in contrast to usual findings with regard to tax evasion by individuals because an increase in activity, say in working time, reduces leisure which, in turn, alters the marginal utility from income (see ▶ "Tax Evasion by Individuals"). Second, the net gain from tax evasion, $Et - (1 - p)C$, is independent of the activity level, x.

   The empirical evidence on the lack of a correlation between firm size and tax evasion activities is mixed (Rice 1992; Nur-tegin 2008; Hanlon et al. 2007; Cai and Liu 2009; Tedds 2010; Hoopes et al. 2012). Since firm size in these studies is measured by the number of employees, value added, or assets, these measures are imperfect proxies for economic activity. Moreover, the costs of evasion may vary with firm size. Accordingly, the empirical findings do not necessarily provide comprehensive evidence with regard to the separability prediction.
3. A higher tax rate, t, will raise evasion activities as long as a rise in t does not increase the marginal costs of evasion, C_E, by more than $1/p$. This will, for example, always be the case if the costs of evasion depend neither on the tax rate ($C_t = 0$) nor vary with the amount of taxes evaded, Et. Accordingly, if adhering to the law becomes more expensive, violations of these regulations will become more severe. This is a straightforward and intuitive prediction which also generally obtains in standard models of the economic analysis of crime but does not always apply in the case of tax evasion by individuals (see Yitzhaki 1974 and the literature cited above in section "Introduction"). The theoretical prediction that higher tax rates induce firms to expand evasion activities is generally consistent with empirical findings (Rice 1992; Nur-tegin 2008; Cai and Liu 2009).
4. If a higher fine, F, raises the marginal costs of tax evasion, so that $C_{EF} > 0$ holds, a rise in F will reduce tax evasion. A greater probability, $1 - p$, of being audited will have the same effect. The empirical evidence, although scarce, is consistent with these predictions (Nur-tegin 2008; Hoopes et al. 2012). Moreover, the separability feature implies that changes in tax enforcement have no impact on a firm's activity level.

5. The results outlined above are not affected by the type of tax considered (Yaniv 1995). Accordingly, the findings are independent of the tax base (such as corporate income, the payroll, value of property, sales, value added), the curvature of the tax schedule, and whether firms act as withholding agents or not. In addition, the theoretical analysis implies that empirical findings for one tax are applicable to others as well.

Robustness

Many analytical contributions of tax evasion by firms have investigated the robustness of the separability result and of its implications and have identified conditions under which the result will no longer hold. In terms of Eq. 1, this can be the case if nonlegal profits, π^N, are affected by a variation in activity, x, not only via official net profits $\pi(x, t)$, i.e., if $Et - (1 - p)C$ varies with activity, x. Such dependence can result under a variety of circumstances, and we discuss four of them below:

1. Assume that $p = p(x)$ holds. (Cf. Virmani 1989). This could be the case, for example, because tax authorities condition the audit probability, $1 - p(x)$, on firm size or because it is related to the frequency of transactions between buyers and sellers which, in turn, are positively correlated to activity. Instead of directly assuming a relationship between p and x, Lee (1998), Marrelli (1984), Marrelli and Martina (1988), and Wang (1990) indirectly include the activity level into the respective specification of the audit probability. Further, Bayer and Cowell (2009) derive such linkage on the basis of a relative auditing rule. According to this rule, the audit probability is positively related to tax payments by other comparable firms and their output levels in an oligopolistic market.

   Given the modification of $p = p(x)$, the first-order conditions for a profit maximum are given by $\pi_x + p'C = 0$ and by (2b). If the audit probability, $1 - p(x)$, declines with activity, because larger firms can avoid detection more easily, tax evasion will induce a firm to expand activity. Fighting tax evasion, for example, by raising the fine, F, and the marginal costs of evasion, C_E, will still reduce evasion activities for a given activity level. Therefore, the gain from an activity expansion will become less pronounced and the overall activity change in response to a higher fine will generally be ambiguous. Even more surprising, a rise in activity, x, enhances the gain, $(1 - p(x))C_E$, from evasion because of the decline in the audit probability. Consequently, the change in evasion activities resulting from a higher fine is also potentially ambiguous (Virmani 1989). Note finally that if knowing the activity level would enable tax authorities to infer the tax base, the assumption of $p = p(x)$ could be inconsistent with the notion of the firm being able to evade taxes. However, knowledge of activity, x, does not necessarily imply that the true tax base is known, unless they are perfectly correlated as, for example, in the case of a unit tax on output.
2. Assume, alternatively, that the firm is not free to choose the under-declaration, E, but can only select a fraction or multiple, e, of activity, x, so that $E = ex$. Such a situation may arise if activity is easily observable, for example, in the case of output levels. The first-order conditions for a profit maximum are given by $\pi_x - e(t - (1 - p)C_E) = 0$ and $t - (1 - p)C_E = 0$. Therefore, activity and evasion decisions are separable if firms can select e optimally. However, if institutional restrictions limit a firm's choice of e, $t - (1 - p)C_E \neq 0$ will hold, and the separability feature will no longer apply (Marrelli 1984; Wang and Conant 1988; Yaniv 1995). Since the restriction on evasion will only be relevant if it is binding, the above result suggests that firms which are easier to monitor and cannot choose e optimally will evade less tax, while evasion and activity choices will be related. The empirical evidence that state-owned companies (Nur-tegin 2008) and publicly traded firms (Rice 1992; Hanlon et al. 2007; Tedds 2010) evade less is consistent with this interpretation.

3. A relationship between economic activity and tax evasion can also arise due to market equilibrium effects. To illustrate, suppose that aggregate activity is positively related to profits because higher profits will induce more firms to enter the market. A higher fine, F, not only raises the costs of evasion ($C_{EF} > 0$) and reduces profits but also makes market entry less attractive. In consequence, the market structure will change along with any policy to reduce tax evasion, and such a policy may actually mitigate competition. Conversely, policies to enhance competition, such as the abolition of product market regulations, may have the detrimental effect of intensifying tax evasion activities by firms (cf. Cai and Liu 2009; Goerke and Runkel 2011).
4. Finally, consider the existence of principal-agent problems. They are likely to arise, for example, in firms run by managers. Managers are unlikely to maximize profits but will pursue their own objectives, at least to some extent. In the presence of such principal-agent problems, the objective function of the firm's decision-maker could be given by $\pi^N(x, E) + K$, where K represents the manager's additional objective. K may be increasing in x and could depend on the under-declaration, E. Moreover, K will vary with the fine and audit probability if the manager is legally or otherwise responsible for tax evasion activities by the firm. Hence, $K = K(x, t, E, F, p)$. Since the manager's first-order condition for a maximum with regard to output is given by $\pi_x + K_x = 0$, it is obvious that activity, x, may vary with tax enforcement and that separability may no longer hold if the decision-maker's marginal payoff is altered by evasion (i.e., if $K_{xE} \neq 0$). As a consequence, changes in the structure of corporate governance; the relative importance of the nonprofit objective, K; restrictions on the level and composition of a manager's remuneration; and the manager's preferences can affect the choice of economic activity, x, and create a relationship, as well as influence the possible linkage between activity and tax evasion activities (see, e.g., Joulfaian 2000; Crocker and Slemrod 2005; Goerke 2007).

Future Directions

In lieu of a summary, we may speculate about future areas of research. For example, the interaction of evasion and avoidance behavior of individuals on the one hand and of firms and businesses on the other can play a greater role in the future. Moreover, the relationship between the structure of input and output markets and tax evasion by firms which are active on these markets is likely to become a more prominent topic. Furthermore, the ability of firms to shift activities across jurisdictions can affect tax evasion and avoidance behavior. In addition, firms are likely to have better outside options than individuals in negotiations with tax authorities. Such bargains between tax payers and authorities also appear to be a promising area of future work. Additionally, the consequences of the division of labor have been an issue which has hardly been looked at. The tax declaration may, for example, be decided upon by a different agent within the firm or outside its realm (think of tax preparers, etc.) than the one who determines the activity level. Such a separation of responsibilities will create additional principal-agent problems. When looking at individuals, recent years have seen intensified attempts to apply various facets of behavioral economics to the analysis of tax evasion (Hashimzade et al. 2013). This may also be an aspect which becomes relevant to the investigation of firm behavior (Alm and McClellan 2012). Finally, it is noteworthy that the economic analysis of crime focuses very much on normative issues. The question of how much tax evasion is optimal has not been an issue looked at intensively with reference to taxes remitted by firms. Hence, it can be conjectured that welfare issues will also play a more prominent role in the future.

References

Alm J (1999) Tax compliance and tax administration. In: Hildreth WB, Richardson JA (eds) Handbook on taxation. Marcel Dekker, New York, pp 741–768

Alm J, McClellan C (2012) Tax morale and tax compliance from the firm's perspective. Kyklos 65(1):1–17

Andreoni J, Erard B, Feinstein J (1998) Tax compliance. J Econ Lit 36(2):818–860
Bayer R, Cowell FA (2009) Tax compliance and firms' strategic interdependence. J Public Econ 93(11–12):1131–1143
Cai H, Liu Q (2009) Competition and corporate tax avoidance: evidence from Chinese industrial firms. Econ J 119(537):764–795
Cowell FA (2004) Carrots and sticks in enforcement. In: Aaron H, Slemrod J (eds) The crisis in tax administration. Brookings Institution Press, Washington, DC, pp 230–275
Cremer H, Gahvari F (1992) Tax evasion and the structure of indirect taxes and audit probabilities. Public Financ/Financ Publiques 47(Suppl):351–365
Crocker KJ, Slemrod J (2005) Corporate tax evasion with agency costs. J Public Econ 89(9–10):1593–1610
Franzoni LA (2009) Tax evasion and avoidance. In: Garoupa N (ed) Encyclopedia of law and economics. Criminal law and economics, vol 3. Edward Elgar, Cheltenham/Northampton, pp 290–319
Goerke L (2007) Corporate and personal income tax declarations. Int Tax Public Financ 14(3):281–292
Goerke L, Runkel M (2011) Tax evasion and competition. Scott J Polit Econ 58(5):711–736
Hanlon M, Mills L, Slemrod J (2007) An empirical examination of corporate tax noncompliance. In: Auerbach AJ, Hines JR, Slemrod J (eds) Taxing corporate income in the 21st century. Cambridge University Press, Cambridge, UK, pp 171–210
Hashimzade N, Myles GD, Tran-Nam B (2013) Applications of behavioural economics to tax evasion. J Econ Surv 27(5):941–977
Hoopes JL, Mescall D, Pittmann JA (2012) Do IRS audits deter corporate tax avoidance. Account Rev 87(5):1603–1639
Joulfaian D (2000) Corporate income tax evasion and managerial preferences. Rev Econ Stat 82(4):698–701
Keen M (2013) The anatomy of the VAT. Natl Tax J 66(2): 423-446
Lee K (1998) Tax evasion, monopoly, and nonneutral profit taxes. Natl Tax J 51(2):333–338
Marchese C (2004) Taxation, black markets, and other unintended consequences, Chapter 10. In: Backhaus JG, Wagner RE (eds) Handbook of public finance, vol 1. Kluwer Academic, Boston, pp 237–275
Marrelli M (1984) On indirect tax evasion. J Public Econ 25(1–2):181–196
Marrelli M, Martina R (1988) Tax evasion and strategic behaviour of the firms. J Public Econ 37(1):55–69
Nur-tegin KD (2008) Determinants of business tax compliance. B E J Econ Anal Policy 8(1), (Topics), Article 18
OECD (2013) Revenue statistics 2013. OECD Publishing
Rice EM (1992) The corporate tax gap: evidence on tax compliance by small corporations. In: Slemrod J (ed) Why people pay taxes – tax compliance and enforcement. The University of Michigan Press, Ann Arbor, pp 125–161
Sandmo A (2012) An evasive topic: theorizing about the hidden economy. Int Tax Public Financ 19(1):5–24
Slemrod J (2007) Cheating ourselves: the economics of tax evasion. J Econ Perspect 21(1):25–48
Slemrod J, Yitzhaki S (2002) Tax avoidance, evasion, and administration, Chapter 22. In: Auerbach AJ, Feldstein M (eds) Handbook of public economics, vol 3. Elsevier, Amsterdam, pp 1423–1470
Tedds LM (2010) Keeping it off the books: an empirical investigation into the characteristics of firms that engage in tax non-compliance. Appl Econ 42(19):2459–2473
Virmani A (1989) Indirect tax evasion and production efficiency. J Public Econ 39(2):223–237
Wang LFS (1990) Tax evasion and monopoly output decisions with endogenous probability of detection. Public Financ Q 18(4):480–487
Wang LFS, Conant JL (1988) Corporate tax evasion and output decisions of the uncertain monopolist. Natl Tax J 41(4):579–581
Yaniv G (1995) A note on the tax evading firm. Natl Tax J 48(1):113–120
Yitzhaki S (1974) A note on income tax evasion: a theoretical analysis. J Public Econ 3(2):201–202

Tax Evasion by Individuals

Laszlo Goerke
IAAEU (Institute of Labour Law and Industrial Relations in the European Union),
University Trier, Trier, Germany
IZA, Bonn, Germany
CESifo, Munich, Germany

Abstract

The basic deterrence model of tax evasion is described, its main predictions are derived, and limitations and flexibility are outlined. Further, the model is interpreted in light of some key institutional features characterizing tax enforcement in OECD countries. Throughout the survey, findings originating from the deterrence model are contrasted with predictions which result from a simple model of criminal activity and law enforcement.

Definition

Tax evasion by individuals represents the attempt to illegally reduce the payment of taxes which

have to be remitted by an individual tax payer to below the level prescribed by law.

Introduction

The analysis of income tax evasion by economists has covered many issues, and most extant surveys focus on a selection of relevant aspects (see, e.g., the contributions by Andreoni et al. (1998), Alm (1999, 2012), Cowell (2004), Slemrod and Yitzhaki (2002), Marchese (2004), Slemrod (2007), Franzoni (2009), and Sandmo (2012)). In many of these reviews, the investigation of tax evasion is interpreted as a special case of the approach which is employed in the economic analysis of crime. In this survey, we explicitly adopt such a perspective and relate findings originating from the analysis of income tax evasion to the broader economics literature on crime. In doing so, we first take a theoretical perspective, present the basic deterrence model of tax evasion, derive its main predictions, and indicate its restrictions as well as the analytical flexibility. Second, we adopt a more institutional viewpoint and confront the theoretical predictions with basic features of real-world enforcement systems. Finally, we compare selected aspects which are discussed in both the literature on tax evasion and the public enforcement of law (as reference for the literature on the public enforcement of law, we use the article in the *Handbook of Law and Economics* (Polinsky and Shavell 2007)).

Basic Theory

We consider a representative, risk-averse individual who is endowed with an exogenously given income Y. This income represents the tax basis and is subject to a linear tax at the rate t. The individual can decide on the amount of income X he/she does not report to tax authorities. Therefore, the gain from evading taxes will equal Xt if tax evasion remains undetected. This takes place with an exogenously given probability p, $0 < p < 1$, and the individual's income then amounts to $Y^s = Y(1-t) + Xt$. With the opposite probability, $1 - p$, the individual or taxpayer will be audited, and tax evasion will be detected. In this case, a fine F is imposed, and the resulting income equals $Y^c = Y^s - F$. While the fine is assumed to be a function of undeclared income X in the seminal contribution by Allingham and Sandmo (1972), Yitzhaki (1974) argues that the penalty is usually based on the amount of taxes evaded, Xt. Consequently, we define the fine F as a linear combination of both determinants, $F := fX[\alpha t + (1-\alpha)]$, where f, $f > 0$, is labeled marginal fine. The parameter α, $0 \leq \alpha \leq 1$, depicts the relative importance of the amount of taxes evaded. For $\alpha = 1$ (0), this implies that the fine is solely a function of taxes evaded (undeclared income). The specification of F reflects the fact that the penalty rates in many OECD countries vary with amount of undeclared taxes but include fixed components or change with other determinants than the underdeclaration (OECD 2009, p. 136 ff). As the final building block, we assume that utility u is increasing in disposable income at a decreasing rate, $u' > 0 > u''$, and that the individual can be described by von Neumann-Morgenstern preferences. Accordingly, expected utility U(X) is given by

$$\begin{aligned} U(X) = {} & pu\underbrace{(Y(1-t) + Xt)}_{:=Y^s} \\ & + (1-p)u\underbrace{(Y(1-t) + Xt - fX(\alpha t + 1 - \alpha))}_{:=Y^c} \end{aligned} \tag{1}$$

Maximizing U with respect to the underdeclaration, X, yields as first-order condition

$$\begin{aligned} U'(X) = {} & pu'(Y^s)t \\ & + (1-p)u'(Y^c)(t - f(\alpha t + 1 - \alpha)) = 0 \end{aligned} \tag{2}$$

The first term in Eq. 2 describes the utility gain from underdeclaring an extra unit of income if tax evasion is successful, while the second term depicts the loss because income declines when being punished. The underdeclaration which results when these two effects are balanced out

is indicated by X*. Note that there will only be an underdeclaration if the gain from evading the first euro of taxes is positive, that is, if $U'(X)$ is greater than zero for $X = 0$ and, hence, for $Y^s = Y^c$. This implies that there is an upper level for the marginal fine $f_{max} = t/[(1 - p)(\alpha t + 1 - \alpha)]$. Furthermore, tax evasion will only be costly if disposable income Y^c shrinks with the underdeclaration in the case of detection. Accordingly, there is a minimal marginal fine $f_{min} = (1 - p)\, f_{max} = t/(\alpha t + 1 - \alpha) > t$. The setting described above focuses on tax evasion. While evasion is the illegal attempt to reduce tax payments, tax avoidance is often interpreted as its legal counterpart. By setting the detection probability, $1 - p$, equal to unity and adding a cost function which increases in the amount of taxes avoided at an increasing rate, the above framework can be amended in order to analyze tax avoidance. Furthermore, many findings derived with regard to tax evasion also hold in an avoidance setting.

Central Results

How does the optimal underdeclaration, $X^* > 0$, vary with income, Y, the parameters of the tax enforcement system, p and f, and the tax rate, t? The respective effects often depend on whether the fine, F, varies with the tax rate, t, i.e., on the value of the parameter α and on the relationship between income and the Arrow-Pratt measure of absolute risk aversion, $R_a(Y) := -u''(Y)/u'(Y)$.

Income, Y, exerts a positive impact on the optimal underdeclaration, X*, if the individual exhibits decreasing absolute risk aversion, R_a, that is, if the willingness to engage in risky activities rises with income. To provide an intuition, note that a higher exogenous income, Y, raises disposable income for a given underdeclaration, irrespective of whether tax evasion is detected or not. If this general increase in income makes the individual more willing to take risks, the gain from higher income shrinks by less than the costs in terms of utility. Therefore, the optimal underdeclaration, X*, rises. Since the tax basis, Y, becomes larger, the amount of taxes paid, that is, $(Y - X^*)t$, may nevertheless increase. If, however, absolute risk aversion, R_a, does not vary with income, there is no income effect, and the underdeclaration remains constant, while the amount of taxes paid surely increases.

A higher marginal fine, f, and a greater detection probability, $(1 - p)$, both reduce the optimal underdeclaration, X*. If the marginal fine, f, rises, two consequences strengthen the incentives to pay taxes. First, there is an income effect since a higher fine payment decreases disposable income, Y^c, if evasion is detected. Therefore, the marginal utility of income in this state of the world rises, and the utility loss resulting from the fall in income if penalized becomes larger. Second, the penalty on the last euro of underdeclared income rises. A higher probability of detection, $1 - p$, makes it more likely that an income loss occurs. Consequently, the individual responds by reducing the loss in disposable income if this more likely event takes place.

The consequences of a higher tax rate, t, hinge on the specification of the fine and on absolute risk aversion, R_a. Suppose, initially, that the fine, F, depends on the amount of taxes evaded ($\alpha = 1$). The optimal underdeclaration, X*, will decline with the tax rate, t, if the individual exhibits constant or decreasing absolute risk aversion, R_a, while the impact is theoretically ambiguous otherwise. For $\alpha = 1$, F is a multiple of the tax rate, t. Accordingly, a rise in t alters the gain and costs from evasion proportionately. Therefore, the impact of the tax rate, t, on the optimal underdeclaration, X*, is solely determined by the income effect. A higher tax rate, t, reduces disposable income and does so more if evasion is detected than if it remains unobserved. If a decline in income, in turn, raises absolute risk aversion, the optimal underdeclaration, X*, will shrink. If, alternatively, the fine, F, depends on the underdeclaration ($\alpha = 0$), X* will rise with the tax rate, t, if absolute risk aversion, R_a, is constant or increasing with income, while the relationship will once again be ambiguous otherwise. In this case, a higher tax rate, t, reduces the penalty relative to the gain from evasion, namely, the lower tax payment. A relative decline in the penalty induces the individual to raise the underdeclaration, X*, ceteris paribus. This substitution effect will be mitigated or reversed by the

income effect which provides greater incentives to underdeclare if absolute risk aversion is declining with income.

While the above analysis has assumed a representative individual, one can easily incorporate heterogeneous taxpayers, for example, in terms of gross income, Y; the degree of absolute risk aversion, R_a; or the marginal tax rate, t. Thus, the analytical model can be used to predict that individuals facing a higher marginal tax rate, t, are more likely to evade and not to pay any taxes, since the maximal and the minimal fines f_{max} and f_{min} increase with t for $\alpha < 1$.

Relating the predictions derived above to the findings obtained in the analysis of crime, it may be observed that the model of criminal activity often employed is based on the assumption that a crime is either committed or not, while the extent of criminal activity per individual is constant. Higher fines and a greater detection probability reduce the incentives to undertake criminal actions, while a higher potential gain will raise them. The latter prediction may be compared to a change in the tax rate, t, derived above. The impact of a higher gross income, or of wealth, depends on whether income also increases disposable income when the criminal is penalized, inter alia. However, the degree of risk aversion does not play a role in the basic setup. These partial differences with respect to the effect of changes in exogenous parameters indicate that predictions can depend crucially on the underlying view of the illegal activity. Does it represent a simple portfolio choice with one safe and another risky asset, as in the case of tax evasion, or does it constitute an endeavor which can be separated from other income-generating activities?

Extensions

The basic deterrence model of income tax evasion has been expanded in numerous ways. We subsequently sketch two extensions which further clarify the sensitivity of the predictions but also the flexibility of the analytical approach. First, we incorporate the idea that individuals will generally be able to decide on the amount of gross income they earn. This decision is likely to result from a trade-off between higher disposable income on the one hand and a greater disutility from generating this income on the other. Hence, utility may be given by $U(Y^s, Y)$ and $U(Y^c, Y)$, depending on whether evasion is detected or not. In addition, $\partial U/\partial Y < 0$ captures the disutility of generating income. In an early contribution, Pencavel (1979) showed that virtually all predictions developed above will not necessarily hold in such a setting. The reason is that any activity which makes tax evasion less attractive also reduces the incentives to generate income. This reduction in the tax base, in turn, lowers evasion activities for a given underdeclaration and, hence, strengthens the incentives to evade. The net effect is generally uncertain because of the differential changes in the marginal utility of disposable income (i.e., $\partial U/\partial Y^s$ and $\partial U/\partial Y^c$) and from generating Y (i.e., $\partial U/\partial Y$). This first extension is an impressive example of the sensitivity of predictions with regard to incorporating additional choice variables.

The individual considered above has occasionally been termed an "amoral tax payer" (Crocker and Slemrod 2005, p. 1595), because the tax evasion decision results solely from the comparison of monetary gains and losses. Therefore, secondly, the question arises how the optimal underdeclaration will be affected if there is a norm with regard to paying taxes. In a simple extension of the basic model, it can be presumed that tax evasion imposes a utility loss on individuals who evade taxes. It has, inter alia, been assumed that this loss (1) is constant, (2) increases in the extent of individual tax evasion, (3) depends on how tax revenues are spent, or (4) varies with an aggregate measure of tax evasion (see Alm and Torgler 2011). While the existence of a norm imposes additional costs of tax evasion in cases (1) and (2) and, therefore, mitigates such activities, the impact in cases (3) and (4) is less obvious. This can be illustrated by assuming that the utility loss from violating the norm of paying taxes varies across individuals and becomes weaker the more people evade taxes. Then, the model may have (at least) two equilibria. In the first, many or all individuals evade taxes, and the

norm does not really bite. In the other equilibrium, few individuals evade taxes. Therefore, the norm imposes substantial costs of evasion, and this helps to stabilize the equilibrium with few people underdeclaring income. In such a setting, the impact of changes in exogenous parameters can be reversed. To illustrate, suppose that higher fines weaken the societal norm of paying taxes. In this case, more severe penalties will reduce evasion, ceteris paribus, but weaken the norm and may induce a move from a low-evasion to a high-evasion equilibrium. In this case, the standard prediction that higher fines reduce illegal activities may no longer hold. This second extension clarifies that the standard deterrence model of tax evasion is flexible enough to be applicable to taxpayers whose preferences include nonmonetary components such as norms.

An Institutional Perspective

Given the importance of the assumptions underlying the model presented above, it is instructive to view them in light of essential features characterizing real-world tax and enforcement systems. In most OECD countries, the nominal income tax rises with income, suggesting that the tax system is progressive. Moreover, the marginal tax burden on wages, taking into account exemptions and government benefits, also generally rises with income, although the marginal rate may decline at specific income levels (OECD 2013). Therefore, the tax rate depends on gross income, $t = t(Y)$, or declared income $Y - X$ in the case of successful evasion. Since it would be optimal to overdeclare income if the marginal tax rate is negative, increasing marginal tax rates have usually been analyzed. While the impact of changes in the enforcement system is generally unaffected by the nature of the tax system, the consequences of changing the marginal tax rate or the progressivity of the tax schedule can also depend on what individuals decide on, namely, the magnitude of the underdeclaration (as in this setting) or of voluntary tax payments (cf., e.g., Yitzhaki 1987; Goerke 2003). Consequently, the tax schedule on its own may affect tax evasion.

The penalty rates in many OECD countries are considerably lower than 100%, even for severe cases of tax evasion (OECD 2009, p. 136 ff). Additionally, the cursory evidence available suggests that the detection probability with regard to income tax evasion is perhaps as low as 1% (see Slemrod 2007 for corresponding information for the United States). In order to integrate this information into the analytical setup, suppose that the utility function u is given by $u(Y) = Y^{(1-a)}/(1 - a)$ and hence features constant relative risk aversion, $R_r(Y) := -u''(Y)Y/u'(Y) = a$. Moreover, the fine is a function of the amount of taxes evaded ($\alpha = 1$); the marginal fine, f, equals 2; the detection probability is assumed to be 10% ($p = 0.9$); and the tax rate is set to $t = 1/3$. Substituting these values into the basic model (cf. Eqs. 1 and 2), the fraction of gross income, Y, which is optimally underdeclared will only be less than 100% if relative risk aversion, R_r, exceeds two. Moreover, the optimal underdeclaration shrinks with R_r, given the above specification of the utility function. If, for example, a value of $R_r = 10$ is assumed, the individual would still underdeclare about 22% of the gross income. Therefore, it has been argued that the standard model seriously overpredicts tax evasion for plausible values of relative risk aversion, R_r, such as between one and five, given the parameters of the tax enforcement system observed in most countries (cf. Alm et al. 1992; Feld and Frey 2002, inter alia).

The response to this criticism has been manifold: Firstly, it has been argued that the payoff of taxpayers is not only affected by the monetary gains and cost of evasion activities but also by the gain of adhering to, or the cost of violating, a social norm, as outlined above. Moreover, the gain may be altered, for example, by whether taxpayers can decide on and approve of the use of tax revenues or how they perceive tax authorities. Secondly, the use of alternative specifications of preferences has been suggested, such as rank-dependent expected utility or prospect theory (cf. Alm and Torgler 2011; Hashimzade et al. 2013). Thirdly, it has been maintained that the numerical example provided above is not an appropriate one with regard to the decision of wage earners but only with respect to self-employed or small businesses (Slemrod 2007).

Wage income is generally subject to withholding regulations. Therefore, the probability that evasion of such income will be detected may approach 100%. A detection probability of 50%, however, would eradicate all evasion incentives in the above numerical example, and the deterrence model of tax evasion can, thus, be reconciled with the data.

Tax Evasion and the Economics of Crime

As mentioned at the outset, the investigation of tax evasion is often interpreted as an application of the economic analysis of crime. However, the perspectives of the two approaches are fundamentally different. A large majority of contributions on tax evasion ask either positive or incrementally normative questions, such as how the tax structure affects evasion activities or whether a certain tax structure is to be preferred to another (see, e.g., the survey by Slemrod and Yitzhaki (2002), in which only one (long) out of eight sections deals with normative issues). The economic analysis of crime focuses strongly on the enforcement of legal rules by public institutions, and the "general problem of public law enforcement may be viewed as one of maximizing social welfare" (Polinsky and Shavell 2007, p. 406). A basic result of this approach is that in the presence of risk-neutral individuals and monetary fines, which have no direct welfare effects, the optimal expected monetary fine should equal the harm caused by a crime. If the sanction is nonmonetary, such as a prison sentence, the expected penalty should be lower because a nonmonetary sanction increases enforcement costs and, thus, lowers welfare.

In the analysis of tax evasion, however, such normative issues have played a comparatively minor role. In one important exception, Slemrod and Yitzhaki (1987) inquire as to what the optimal size of a tax collection agency is in the presence of risk-averse individuals. For a given amount of tax revenues, less tax evasion mitigates income variability, and this reduction in the "excess burden of tax evasion" (Slemrod and Yitzhaki 1987, p. 187) represents the welfare gain from reducing evasion activities. Higher enforcement costs constitute the welfare loss due to fighting tax evasion. The optimal degree of law enforcement is attained when the revenue effect of stricter enforcement still exceeds the resource costs of achieving this revenue impact. The reason is that the revenue gain is mainly distributionary and has no direct welfare impact in a setting with identical individuals, while costs of enforcement reduce welfare. This finding resembles those obtained in the economic analysis of crime.

Given the different perspectives, the literature on tax evasion and the contributions on public law enforcement have also approached many extensions of the basic settings in alternative ways. To illustrate, we consider the nature of penalties and settlements.

From an economics of crime perspective, nonmonetary sanctions, such as prison sentences, can have two advantages over fines. Firstly, the financial means of a tax evader may be insufficient to pay a fine. However, imprisonment is feasible irrespective of wealth so that nonmonetary penalties may still deter illegal activities when monetary fines no longer have this effect. Secondly, imprisonment generally limits future crimes. Such an incapacitation effect is less likely to occur in the case of monetary penalties. One important disadvantage of nonmonetary penalties is the higher cost of enforcing such penalties. In most countries, the penalties for evading personal income taxes are represented by monetary fines. However, for severe cases of tax evasion, prison sentences can also be imposed (cf. OECD 2009, Table 31). Nonetheless, questions such as (1) what are the effects of monetary and nonmonetary penalties on tax evasion?, (2) when should imprisonment be used and sentences be suspended?, and (3) what is the optimal combination of fines and imprisonment? have not figured prominently in contributions on tax evasion. This is in contrast to the literature on public law enforcement.

Settlements, that is, agreements between an offender and authorities to terminate or avoid a court trial in exchange for accepting a penalty, have received substantial attention in the economic analysis of crime. Settlements can be desirable because they reduce the costs of law enforcement. Furthermore, risk-averse individuals may prefer

certain penalties to uncertain court outcomes. The main disadvantage of settlements is that they will be attractive to offenders only if they effectively imply a lower penalty. This dilutes the deterrence effect of sanctions. In addition, a settlement may hinder the detection of all illegal activities of an offender and can prevent the development of precedents. While such aspects of settlements have been discussed in the literature on the public enforcement of law (Polinsky and Shavell 2007, p. 435 f), there are few relevant contributions relating to tax evasion (Macho-Stadler and Pérez-Castrillo 2004; Franzoni 2004).

The relative infrequency of settlements may be due to the fact that trials in cases of tax evasion are much less frequent than for criminal activities such as theft, fraud, physical injury, or murder. However, the perspective can also be reversed. Often, tax authorities impose a penalty. This procedure may be interpreted as the tax authority's (pretrial) proposal of a settlement. Accordingly, the relevant question in the context of tax evasion may not be whether settlements are beneficial but why they are used so extensively.

The two above examples clarify that the investigation of topics analyzed in the public enforcement of law may also generate additional insights in the context of income tax evasion. Other such issues may relate to the self-reporting of past tax evasion activities, the treatment of repeat offenders, the employment of tax advisors, corruption among enforcement agents, and the role of marginal deterrence. The analysis of such topics will be especially rewarding if institutional features of tax evasion activities are taken into account. Such investigations would help to clarify whether or not predictions based on general models of illegal behavior carry over to the more specific settings applicable to the investigation of income tax evasion.

References

Allingham MG, Sandmo A (1972) Income tax evasion: a theoretical analysis. J Public Econ 1(3–4):323–338

Alm J (1999) Tax compliance and tax administration. In: Hildreth WB, Richardson JA (eds) Handbook on taxation. Marcel Dekker, New York, pp 741–768

Alm J (2012) Measuring, explaining, and controlling tax evasion: lessons from theory, experiments and field studies. Int Tax Public Financ 19(1):54–77

Alm J, Torgler B (2011) Do ethics matter? Tax compliance and morality. J Bus Ethics 101(4):635–651

Alm J, McClelland GH, Schulze WD (1992) Why do people pay taxes? J Public Econ 48(1):21–39

Andreoni J, Erard B, Feinstein J (1998) Tax compliance. J Econ Lit 36(2):818–860

Cowell FA (2004) Carrots and sticks in enforcement. In: Aaron H, Slemrod J (eds) The crisis in tax administration. Brookings Institution Press, Washington, DC, pp 230–275

Crocker KJ, Slemrod J (2005) Corporate tax evasion with agency costs. J Public Econ 89(9–10):1593–1610

Feld LP, Frey BS (2002) Trust breeds trust: how taxpayers are treated. Econ Gov 3(2):87–99

Franzoni LA (2004) Discretion in tax enforcement. Economica 71(283):369–389

Franzoni LA (2009) Tax evasion and avoidance. In: Garoupa N (ed) Criminal law and economics, vol 3, 2nd edn, Encyclopedia of law and economics. Edward Elgar, Cheltenham/Northampton, pp 290–319

Goerke L (2003) Tax evasion and tax progressivity. Public Financ Rev 31(2):189–203

Hashimzade N, Myles GD, Tran-Nam B (2013) Applications of behavioural economics to tax evasion. J Econ Surv 27(5):941–977

Macho-Stadler I, Pérez-Castrillo D (2004) Settlement in tax evasion prosecution. Economica 71(283): 349–368

Marchese C (2004) Taxation, black markets, and other unintended consequences, Chapter 10. In: Backhaus JG, Wagner RE (eds) Handbook of public finance, vol 1. Kluwer, Boston, pp 237–275

OECD (2009) Tax administration in OECD and selected non-OECD countries: comparative information series (2008). OECD, Paris

OECD (2013) Taxing wages 2013. OECD, Paris

Pencavel JH (1979) A note on income tax evasion, labor supply, and nonlinear tax schedules. J Public Econ 12(1):115–124

Polinsky AM, Shavell S (2007) The theory of public enforcement of law, Chapter 6. In: Polinsky AM, Shavell S (eds) Handbook of law and economics, vol 1. Elsevier, Amsterdam, pp 403–456

Sandmo A (2012) An evasive topic: theorizing about the hidden economy. Int Tax Public Financ 19(1):5–24

Slemrod J (2007) Cheating ourselves: the economics of tax evasion. J Econ Perspect 21(1):25–48

Slemrod J, Yitzhaki S (1987) The optimal size of a tax collection agency. Scand J Econ 89(2):183–192

Slemrod J, Yitzhaki S (2002) Tax avoidance, evasion, and administration, Chapter 22. In: Auerbach AJ, Feldstein M (eds) Handbook of public economics, vol 3. Elsevier, Amsterdam, pp 1423–1470

Yitzhaki S (1974) A note on income tax evasion: a theoretical analysis. J Public Econ 3(2):201–202

Yitzhaki S (1987) On the excess burden of tax evasion. Public Financ Q 15(2):123–137

Tax Structure

▶ Fiscal System

Tax Systems

▶ Fiscal System

Taxation

▶ Fiscal System

Telecommunications

Alden F. Abbott
Heritage Foundation, Washington, DC, USA

Abstract

Telecommunications involves the transmission of information without change in the form or content. Telecommunications networks increase in value as the number of users rises ("network effect") and the rise of the Internet and wireless communications has bestowed huge economic benefits on countries worldwide. The development of telecommunications has been heavily influenced by regulatory regimes. Regulation in the United States has featured efforts to restrain private monopoly power and promote market allocation of spectrum, while European regulation recently has focused on the privatization of former state telecommunications monopolies and the transition to a pan-European regulatory regime. As the United States transitions away from US government stewardship of the Internet and toward a more global form of Internet governance, the International Telecommunication Union is seeking to play a greater role in Internet oversight. Emerging telecommunications policy issues include the privatization of the radio spectrum, "net neutrality" regulation aimed at treating all data traffic equally, and the growth of "cloud computing" and "big data" compilations.

Conceptual Overview

Telecommunications may be broadly defined as "the transmission, between or among points specified by the user, of information of the user's choosing, without change in the form or content of the information as sent and received" (47 U.S. Code § 153 2011). The information transmitted may take the form of data, text, audio, and/or visual materials.

Whether the telegraph, the telephone, or the Internet, the full value of a telecommunications medium may not be realized until there is a network of users that can transfer information among themselves. This addition in value can be attributed to what is known as a "network effect." There is a significant difference, however, between network effects and network externalities. While the term "network effect" refers to the increase in the value of a network that corresponds with the increase in the number of participants in the network, "network externalities" occur when market participants do not fully "internalize" (obtain) that increase in value, for example, the owners of a private network (Liebowitz and Margolis 1998).

The network effect is an example of a positive externality, in which the action of one individual benefits another individual without any mutual agreement to make compensation for that benefit (Easley and Kleinberg 2010). A prime example is the benefit gained from additional participants in a social networking site, which raises the potential for any given individual to network through that site, even though no participant is explicitly compensated for joining (Easley and Kleinberg 2010). The Internet illustrates the network externality of the social media to a grand degree, with over 200 million Americans having broadband access to the Internet (Broadband Fact Sheet Pew Research Center 2013) and an estimated 2.9 billion Internet users worldwide (International

Telecommunication Union Key ICT Data 2005 to 2015), bringing major economic benefits with it. The Internet "alone accounted for 21% of the GDP growth in mature economies from [2006 to 2011]": Brazil, Canada, China, France, Germany, India, Italy, Japan, Korea, Russia, Sweden, the UK, and the United States (Manyika and Roxburgh 2011). According to one study, among such "high-income" economies from 1980 to 2002, "a 10% increase in broadband penetration yielded an additional 1.21 percentage points of GDP growth," while the same level of broadband penetration in "low- and middle-income economies" yielded 1.38 percentage points of GDP growth (International Telecommunications Union Impact of Broadband 2012).

This value of an established network due to an existing pool of participants has led some to worry that an earlier-created inferior product with a preexisting network of participants would win out against a later-arriving but superior product without such a network. If two competing networks produce similar but incompatible products, the product with the greater market share will have an advantage. If the network effect does not diminish, the advantaged network could be seen as a natural monopoly (Liebowitz and Margolis 1998).

However, the network effect is not limitless. If additional participants cease to provide value to the existing network of participants, meaningful competition among networks may be possible. Subjective valuation of the network by individuals may also differ, allowing multiple competing networks to coexist (Liebowitz and Margolis 1998). Similarly, different individuals may value individual participants added to the network differently, which could provide an opening in the market for separate networks to serve different groups based on the actual makeup of the participants (Liebowitz and Margolis 1998).

The United States and Europe possess the most mature telecommunications regulatory regimes. An overview of these systems provides insight on the sorts of problems national governments face as they oversee the development of their telecommunications sectors. Following the overview, this essay briefly describes international telecommunications regulation and emerging policy issues.

Telecommunications Regulation in the United States of America (USA) and the European Union (EU)

In one of the first attempts to regulate telecommunications as a whole, the US Congress passed the Communications Act of 1934, establishing the Federal Communications Commission (FCC) to take responsibility for regulating radio and wire communications, which had previously been dealt with by separate agencies (Communications Act of 1934). Congress amended the law with the Telecommunications Act of 1996, allowing for federal preemption of local regulations that acted as barriers to entry and more competition in the long-distance market and requiring incumbents to allow access to their networks at wholesale prices (Telecommunications Act of 1996). The FCC also has oversight over wireless services and radio and television broadcasting ("What We Do," Federal Communications Commission 2015), which are licensed to use certain portions of the electromagnetic spectrum.

In 1974, the Department of Justice sued under the Sherman Antitrust Act to rein in the dominant US national telecommunications company, in filing suit against AT&T to modify an existing 1949 consent order. The suit alleged that AT&T, the monopoly provider of local telephone service in most parts of the United States, had engaged in various anticompetitive acts to maintain monopoly power in the provision of long-distance telephone service (U.S. v. AT&T 1978). The suit ended in divestiture for AT&T in a modified consent decree in 1983, effectively breaking up the telecom giant (U.S. v. AT&T 1983). Since 1983, additional antitrust actions have been directed at major US telecommunications companies. For example, in 1998, the United States sued under the Clayton Antitrust Act to block the merger of AT&T with TCI (the merger went forward subject to a consent agreement) (U.S. v. AT&T 1999), and Verizon was sued under the Sherman Antitrust Act for conduct that had been found to violate the Telecommunications Act of 1996 (the antitrust suit failed) (Verizon Communications 2004).

The telecommunications regulatory regime in the United States receives policy advice and technical support from the National

Telecommunications and Information Administration (NTIA), established in 1978. The NTIA advises the President and works with executive branch agencies to develop policy on telecommunication and information issues and manages the federal use of the electromagnetic spectrum by administering grants and holding auctions to assign licenses (About NTIA 2015) (the NTIA has worked with the FCC to reassign certain spectrum frequencies from public use to private use). The NTIA also has been indirectly involved in Internet governance (and, in particular, the administration of the Internet Domain Name System) through its administration of a US government contract with the Internet Corporation for Assigned Names and Numbers, or ICANN, and the Internet Assigned Numbers Authority, or IANA (ICANN 2015).

In Europe, telecommunications was mostly provided by state-owned monopolies until the 1970s, when pushes for a smaller role for governments in the telecommunications market began to rise (Bauer 2013). In the 1980s, the European Commission began to promote a vision of a pan-European telecommunications sector (Bauer 2013). The EU successfully liberalized terminal equipment, value-added, and other services and had opened all services up to competition by 1998 (Bauer 2013). By 2012, all but one member, Luxembourg, had at least partially privatized their telecommunications sector (Bauer 2013). Since then, the European Union has moved toward facilitating regulation and standardization of telecommunications throughout EU member states as part of the *Digital Agenda for Europe*. The EU's framework is made up of five directives, the Framework Directive, the Access Directive, the Authorisation Directive, the Universal Service Directive, and the Directive on Privacy and Electronic Communications, and two regulations, the Regulation on Body of European Regulators for Electronic Communications (BEREC) and the Regulation on Roaming on Public Mobile Communications Network (Digital Agenda for Europe Telecoms Rules 2015). In 2012, the European Parliament and Council approved the first Radio Spectrum Policy Programme (RSPP) to set objectives, make recommendations, and establish principles for the administration of the radio spectrum (Digital Agenda for Europe Radio Spectrum Policy Program 2015). The overall aim of these efforts is to move toward a pan-European approach to telecommunications regulation, in place of the nation-specific regulatory regimes that currently exist within the EU.

Other jurisdictions throughout the world employ a variety of regulatory schemes, with the trend being toward provision of telecommunications services through private operators rather than the state (Struzak 2003). The recent fast international growth of mobile wireless telecommunications, which has rapidly spread the availability of telecommunications services to new populations (especially the poor), is another feature that is expanding the telecommunications network effect and, in particular, widespread access to the Internet.

International Telecommunications Regulation

While the United States has historically acted as steward of the Internet, the NTIA has received pressure to move toward a more global model of Internet governance. As a step toward this model, the NTIA asked ICANN to turn over its role in coordinating the Internet's Domain Name System (DNS) to the international community as part of a program of privatization (NTIA Announces Intent to Transition 2014). In so doing, the NTIA, acting consistently with resolutions of the US Senate and House of Representatives (S. Con Res. 50, 112th Cong., 2012), specified that it would not support handing its responsibility to any governmental or intergovernmental. The current contract expires on September 30, 2015. Some commentators have expressed concerns about the implications of this transition for the future of Internet governance (Schaefer et al. 2015), while others support NTIA's initiative (Llansó 2015; Tennenhouse 2014).

The International Telecommunication Union (ITU), founded in 1865 as the International Telegraph Union (International Telecommunication Union History 2015), is now a United Nations Agency focused specifically on information and communication technologies (ICTs). The

organization has proposed to take up the mantle of the NTIA, resolving to "to explore ways and means for greater collaboration and coordination between ITU and relevant organizations involved in the development of IP-based networks and the future Internet, through cooperation agreements, as appropriate, in order to increase the role of ITU in Internet governance so as to ensure maximum benefits to the global community," specifically mentioning ICANN in its 2014 Plenipotentiary Resolution 102 (ITU Plenipotentiary Resolution 2014).

Emerging Policy Issues

Proposals to increase privatization of the radio spectrum and to regulate more heavily the provision of Internet communications are the subject of considerable recent debate. While auctions have been the chosen method of allocating newly privatized spectrum blocs in the United States, the process is not without controversy. For example, after a January 2014 auction for wireless licenses, AT&T accused Dish Network Corporation of intentionally driving up the price of spectrum licenses at auction, arguing that the coordinated bidding strategy involving three entities artificially inflated the perceived demand for the licenses. Dish Network asserted that it fully complied with FCC rules when implementing the bidding strategy, but there have been calls for the FCC to intervene to prevent such behavior (Gryta and Ramachadran 2015).

More generally, some commentators have advocated in favor of "net neutrality," the principle that Internet service providers (ISPs) should treat all data traffic equally, in the regulation of ISPs (Lessig and McChesney 2006). On February 26, 2015, The FCC adopted a rulemaking that would, among other things, classify the Internet as a "public utility" regulated under Title II of the US Telecommunications Act, invoking the cause of net neutrality. In so acting, the FCC opined that in absent regulatory action, the Internet service providers may throttle data speeds based on content or offer high-paying customers prioritization in data traffic (FCC News Release 2015). Opponents voice concern that the regulation of the Internet as a public utility will result in the entrenchment of larger Internet service providers (ISPs) at the expense of smaller providers, a slowdown in Internet speeds (or the rate of increase in speeds), and Internet access rates (Summary of Pai Testimony 2015). The rule is likely to face challenges from opponents, and the future of Internet regulation by the FCC remains uncertain (Hughes 2015). Concerns about imposing "net neutrality" and various other constraints on the provision of Internet service may be expected in other jurisdictions as well.

Ongoing changes in telecommunications infrastructure will also drive policy debates. Commentators have predicted the increased use of cloud services, investment in telecommunications infrastructure, and replacement of telephone lines with broadband in the developed world and with the rise in multi-device telecommunication services (Lopez 2015). The problem of "big data," or collections of information that "had grown so large that the quantity being examined no longer fit the memory that computers use for processing," has been a focus of discussion among telecommunications experts (Pflugfelder 2013; International Telecommunications Union Press Release December 2014).

Conclusion

The importance and diffusion of telecommunications services may be expected to grow apace, with the rapid expansion of wireless services and Internet-related transactions. This development promises to bestow large and growing benefits on producers and consumers worldwide. Nevertheless, questions about the future of Internet governance and regulation in general create some uncertainty as to the manner in which telecommunications (and, in particular, Internet) service provision will grow and evolve.

References

47 U.S. Code § 153 (2011)

Bauer JM (2013) The evolution of the European regulatory framework for electronic communications. Catedra Telefonica, Institut Barcelona D'Estudis Internacionals,

pp 6–7. http://catedratelefonica.ibei.org/wp-content/uploads/2013/06/IBEI-%C2%B7-41.pdf. Accessed 13 Mar 2015

"Broadband Technology Fact Sheet," Pew Research Center (2013). http://www.pewinternet.org/fact-sheets/broadband-technology-fact-sheet. Accessed 2 Mar 2015

Communications Act of 1934, Public Law 73-416

Corwin PS If Stakeholders already control the Internet, why NETmundial and the IANA transition? CircleI-D. Accessed 13 Mar 2015

Digital Agenda for Europe, "Radio Spectrum Policy Program," (2015) https://ec.europa.eu/digital-agenda/node/173. Accessed 2 Mar 2015

Digital Agenda for Europe, "Telecoms Rules," (2015) https://ec.europa.eu/digital-agenda/en/telecoms-rules. Accessed 17 Mar 2015

Easley D, Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected world. Cambridge University Press, pp 509–510. http://www.cs.cornell.edu/home/kleinber/networks-book/networks-book-ch17.pdf. Accessed 2 Mar 2015

FCC News Release, "FCC Adopts Strong, Sustainable Rules to Protect the Open Internet," *Federal Communications Commission*, February 26, 2015. http://www.fcc.gov/document/fcc-adopts-strong-sustainable-rules-protect-open-internet. Accessed 2 Mar 2015

Gryta T, Ramachadran S (2015, February 20) AT&T Says Dish Bidding Practices Skewed Results in Spectrum Auction. The Wall Street Journal. http://www.wsj.com/articles/at-t-says-dish-bidding-practices-skewed-results-in-spectrum-auction-1424454394. Accessed March 2, 2015

Hughes S (2015, February 25) FCC's Net Neutrality Rules Expected to Unleash Court Challenges. The Wall Street Journal. http://www.wsj.com/articles/fccs-net-neutrality-rules-expected-to-unleash-court-challenges-1424919940. Accessed 12 Mar 2015

"ICANN," *National Telecommunications and Information Administration*, 2015, http://www.ntia.doc.gov/category/icann. Accessed 12 Mar 2015

International Telecommunication Union, "History," 2015. http://www.itu.int/en/about/Pages/history.aspx. Accessed 2 Mar 2015

International Telecommunications Union, Impact of Broadband on the Economy: Research to Date and Policy Issues, April 2012, p 4, http://www.itu.int/ITU-D/treg/broadband/ITU-BB-Reports_Impact-of-Broadband-on-the-Economy.pdf. Accessed 12 Mar 2015

International Telecommunications Union, Key ICT Data for the World, 2005 to 2015, www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx. Accessed 2 Mar 2015

International Telecommunications Union Press Release, "ITU Telecom World 2014 Highlights, Innovations, Technologies, and Ideas Shaping Future of ICTs: Interactive Debates and Showcases Focus on Future of Technology and Its Impact on Society," International Telecommunications Union, December 10, 2014, http://www.itu.int/net/pressoffice/press_releases/2014/76.aspx#.VQiL9I7F98E. Accessed 17 Mar 2015

ITU Plenipotentiary Resolution 102 (Rev. Busan 2014), p 4, http://www.itu.int/en/plenipotentiary/2014/Documents/final-acts/pp14-final-acts-en.pdf. Accessed 13 Mar 2015

Lessig, L, McChesney R (2006, June 8) No Tolls on the Internet. The Washington Post. http://www.washingtonpost.com/wp-dyn/content/article/2006/06/07/AR2006060702108.html. Accessed 12 Mar 2015

Liebowitz SJ, Margolis SE (1998) Network Externalities (Effects) The New Palgrave Dictionary of Economics and the Law. http://www.utdallas.edu/~liebowit/palgrave/network.html. Accessed 2 Mar 2015

Llansó E Don't Let Domestic Politics Derail the NTIA Transition Center for Democracy and Technology. https://cdt.org/blog/dont-let-domestic-politics-derail-the-ntia-transition/.Accessed 12 Mar 2015

Lopez M (2015) What mobile cloud means for your business. Fortes. www.forbes.com/sites/marite/lopez/2015/04/06/what-mobile-cloud-means-for-your-business/. Accessed 20 May 2016

Manyika J, Roxburgh C (2011) The Great Transformer: The Impact of the Internet on Economic Growth and Prosperity (New York: McKinsey Global Institute), pp 1–2, http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_great_transformer. Accessed 16 Mar 2015

NTIA Announces Intent to Transition Key Internet Domain Name Functions. National Telecommunications and Information Administration, Office of Public Affairs, March 14, 2014, http://www.ntia.doc.gov/press-release/2014/ntia-announces-intent-transition-key-internet-domain-name-functions. Accessed 2 Mar 2016

"Overview," *International Telecommunication Union,* http://www.itu.int/en/about/Pages/overview.aspx. Accessed 2 Mar 2015

Pflugfelder EH (2013, August) Big Data, Big Questions. Communication Design Quarterly 1.4, p 18

Schaefer BD, Rosensweig PS, Gattuso JL. Time is Running Out: The U.S. Must be Prepared to Renew the ICANN Contract. Heritage Foundation Issue Brief No. 4340, February 3, 2015, http://www.heritage.org/research/reports/2015/02/time-is-running-out-the-us-must-be-prepared-to-renew-the-icann-contract%20-%20_ftn3. Accessed 12 Mar 2015

S.Con.Res.50, 112th Cong, 2d Sess., Sept. 22, 2012, pp 2–3, http://thehill.com/images/stories/blogs/flooraction/jan2012/sconres50.pdf. Accessed 13 Mar 2015

Struzak R (2003, 2–22 February) Introduction to International Radio Regulations. Lectures given at the School on Radio Use for Information And Communication Technology Trieste, p 6, http://www.iaea.org/inis/collection/nclcollectionstore/_public/38/098/38098197.pdf. Accessed 12 Mar 2015

"Summary of Commissioner Pai's Oral Dissent on Internet Regulation," Press Release, Federal Communications Commission, February 26, 2015, http://www.fcc.gov/document/summary-commissioner-pais-oral-dissent-internet-regulation. Accessed 2 Mar 2015

Telecommunications Act of 1996, Public Law 104-104

Tennenhouse D (2014, March 3) "Microsoft Applauds US NTIA's Transition of Key Internet Domain Name Functions," Microsoft on the Issues. http://blogs.microsoft.

com/on-the-issues/2014/03/17/microsoft-applauds-us-ntias-transition-of-key-internet-domain-name-functions/. Accessed 12 Mar 2015
U.S. v. AT&T, 461 F.Supp. 1314, 1318 (1978)
U.S. v. AT&T, 552 F. Supp. 131, 131 (D.D.C. 1983)
U.S. v. AT&T (Complaint), 1999 WL 1211462 (D.D.C. Aug. 23, 1999)
Verizon Communications v. Law Offices of Curtis V. Trinko, LLP, 540 U.S. 398 (2004)
"What We Do," *Federal Communications Commission,* http://www.fcc.gov/what-we-do. Accessed 11 Mar 2015

Further Reading

Haring, J Telecommunications. In: The concise encyclopedia of law and economics (2nd ed.), http://www.econlib.org/library/Enc/Telecommunications.html. Accessed 17 Mar 2015
World Bank and International Telecommunications Union, Telecommunications Regulation Handbook, Colin Blackman and Lara Srivastava eds. (10th Anniversary ed. 2011)

Television Market

Paola Savini
Autoritá per le garanzie nelle comunicazioni – AGCOM, Rome, Italy

Definition

The television market is made up of content producers, broadcasting network operators, packagers and TV service providers, carriers and network providers, as for the supply side, as well as the so-called "audience," for the demand side. It is a complex TWO-SIDED market, mature and strongly regulated. It is characterized by cultural, political, and industrial elements. These different components are deeply rooted within the national television markets but, at the same time, respond to global economic dynamics and supranational laws. The television industry can be divided into three main activities, corresponding to three different economic functions. These are the programs' production, the organization of television programs in a schedule, and signal distribution in the territory. Almost a century after its invention and standardization, television is nowadays evolving again, as the competition provided by newer and different screens (pc, smartphone, tablet) is increasing, thus offering the audience multiple options to access linear and on demand audiovisual contents.

Television Market: An Overview

In 2016, the European Union hosts more than 4,000 television services (European Audiovisual Observatory 2017), ranging from the historical generalist channels to the thematic ones, offering sports, film and series, children's programs, and documentary to millions of habitants. Also many Video on demand service providers operate within the European countries, and hundreds of such services are available everywhere (Ofcom 2016).

The one-century old television is, in fact, still the mass medium *par excellence*, offering entertainment and information worldwide and accounting for about 20% of the not-sleeping hours in many countries (Ofcom 2017). Most countries currently license a high number of television channels which can operate at a national or subnational (local) level, now accessible via different platforms, such as the digital terrestrial television (DTT), the cable, the satellite, or the Internet protocol television (IPTV), giving to the public the chance, never experienced before, to watch a huge variety of audiovisual contents. This audiovisual consumption may now happen through the TV set, but also on mobile devices, personal computers, tablets, or game consoles.

All of these changings are both technology and policy driven, as more than other media, television is characterized by cultural, political, and industrial elements (see entry on ▶ "Media"). These different components are deeply rooted in the national contexts of each television company (hereinafter, also called "broadcasters") but, at the same time, respond to global economic dynamics and supranational laws (Freedman 2008).

Legal Framework

As regards the legal framework, numerous forms of public authorities exercise some kind of regulation over broadcasting activities, at different levels: apart from national legislative Chambers,

which regulate particularly the governance of the Public Service Media (PSM) and the content' level, there are the Communitarian institutions within the EU framework, some governmental administrative authorities, some independent regulatory authorities (see EPRA – European Platform of Regulatory Authorities – 52 authorities from 46 countries are members – which cooperates with the European Commission, the Council of Europe, the European Audiovisual Observatory and the Office of the OSCE Representative on Freedom of the Media), as well as the national Courts for specific issues. Within some Member State, also some forms of self-regulation apply, especially for Public Service Media.

Media ownership and concentration issues have traditionally been regulated at national level, with regard to media pluralism and freedom of expression in particular. However, some common principles may be found within the "*EU Charter of Fundamental Rights*" (Article 11) and within the "*European Convention for the Protection of Human Rights and Fundamental Freedoms*" (Article 10).

Differences are huge between European countries, as concerns media ownership: in France, some limitations apply both to the ownership of a single television channel and to the ownership of the number of licenses; in Italy, limitations apply both to the cross-ownership between national and local terrestrial channels and to the revenues within the media sector; in Germany, audience market share limits apply.

Concerning other issues such as programming, advertising, sponsorship, and the protection of certain individual rights, some Council of Europe countries signed already in 1989 the "European Convention on Transfrontier Television" (ETS n.132), which created a legal framework for the free circulation of transfrontier television programs in Europe. In particular, the Convention defined for the first time "audiovisual works of European origin" as the "creative works whose production or co-production is controlled by natural or legal European producers."

Then, at EU level and with the neighboring countries, in 1989 it became effective a coordination of national legislations on all audiovisual media, through the so-called "*Television Without Frontiers Directive*" (Directive 89/552/EEC), which aimed to safeguard some public interest objectives, such as cultural diversity, the protection of minors, and the right of reply, and rested on two basic principles: the free movement of European television programs within the internal market and the requirement for TV channels to reserve, whenever possible, more than half of their transmission time for European works (the so-called "broadcasting quotas").

Having been this Directive substantially amended several times during the years (it is currently again under revision, and a new legislative proposal has been adopted by the European Commission on 25 May 2016), a minimum set of common rules applying to all the 28 Member States, it is now operating for the audiovisual and new media sector in the Union, based on the country of origin principle, ensuring freedom of reception and retransmission of the TV channels as well as advertising regulation, promotion of European works, and protection of minors (AVMSD 2010). All audiovisual media services providers (both television broadcasts and contents selected by viewers on-demand over an electronic communication network) shall display clearly their name and address and need to respect the basic rules provided in the AVMSD, as regards: the prohibition of incitement to hatred based on race, sex, religion, or nationality; the accessibility for people with visual or hearing disabilities; some qualitative and quantitative requirements for commercial communications.

At an even more international level, global radio spectrum and satellite orbits as well as technical standards are regulated by the International Telecommunication Union (ITU), the United Nations specialized agency for information and communication technologies.

Industrial Innovations

As regards the industry level, since the 1970s, technological innovations strongly fostered the enlargement of the television market, generally born as state-owned, and enriched the consumers' experience.

In 1962 the first transatlantic transmission, via the Telstar communications satellite, was made, permitting satellite images to be taken from around the globe. The first videotape recorders were sold to the public in 1965. Color broadcasting began in the USA in 1964. In 1972, the first digital converter changed pictures from the US 525-line format into the 625-line European standard. In the same year, the first cable operators appeared in the USA. An audience measurement system was implemented in the USA in 1973 and the UK in 1977 (the British Audience Research Bureau was created in 1981).

At the beginning of 2000, digitalization burst onto the scene as a disruptive innovation that would be able to revolutionize the entire TV ecosystem once again, at various levels of the TV supply chain. This technological innovation has been institutionally recognized as pro-competitive and thus as having positive direct and indirect influences on media pluralism, as a result (Adda and Ottaviani 2005). For the past 30 years the replacement of analog technologies with digital ones has indeed changed the transport systems of the signal: a long process of redefinition for transmission standards has been set up more recently (2009–2015), for the final switch off of the analogue signal, in the OECD countries. Digital terrestrial television was launched in the UK in 1998, just after digital satellite television. In the United States, the cutoff date for analogue TV was established in 2009, in Germany in 2008 and in Italy and UK in 2012.

Compression systems have also facilitated an increase in the transmission capacity of the networks, resulting in the multiplication of channels and storage, production, and distribution at decreasing costs, thus removing a major bottleneck from the whole TV industry chain. Since the cultural and social benefits that have historically been found in the broadcasting activity, free or low cost access to the spectrum has been granted to the broadcasters.

Digital competition has also revitalized the electronics market through the sale of digital receivers, set-top-boxes, televisions HD ready, as well as the 3D TVs or connected TVs.

Furthermore, as regards consumption habits, traditional live viewing is slowly declining everywhere while time-shifted and nonbroadcast viewing activities (such as Video on demand services or streaming video) are rising, through services delivered over the Internet, such as Netflix, Amazon Prime Video, Hulu, iQIYI, Youku, Ditto TV (now absorbed by Zee5), Now TV, with different percentages of subscription within the European countries and the USA (Ofcom 2017).

Resources and Business Models: Advertising, Public Funds, and Subscriptions

The television industry can be divided into three main activities, corresponding to three different economic functions. These are the programs' production, the organization of television programs in a schedule, and signal distribution in the territory (Owen and Wildman 1992) The signal distribution is undertaken by the presence of economic agents who make available to broadcasters the transmission equipment (carriers or network providers). Each distribution channel – operating through analog or digital terrestrial network, cable, satellite – involves different players (manufacturers, device producers, multiple-system operators, regulatory authorities). If the television service has controlled access (as for instance, pay-TV), packagers and pay-TV service providers will also be involved in some economic activities, such as customer relationships or channels bundling.

The cultural industries of most European and other countries have, historically, focused political attention and industrial policies on the scheduling and distribution aspects, as broadcasting systems were generally born as state-owned ones, while the US cultural industry gave centrality to the productive function and to advertising-based, privately owned, broadcasting models. One fundamental component of the television ecosystem is in fact the content, whose characteristics are important to determine the audience a broadcaster or a service provider wants to achieve. The production function can be

concentrated within the same broadcasters, either directly or through in-house production unites/subsidiaries, or it may result from independent firms (producers). In order to qualify as an "independent producer," in Europe a production company must be a company (or other business entity) which is independent of broadcasters. The historical origins of European Union policy on self-sufficient audiovisual production and TV distribution found a cornerstone with the already mentioned "Television without Frontiers Directive." Similar objectives can be found also in some American audiovisual policies, such as the "Prime Time Access Rule" (PTAR) and the "Financial Interest and Syndication Rules" (also known as "Fin–Syn").

As regards the economic functions, the central activity of the television industry is the exchange of "television communication," supplied by broadcasting networks and requested by viewers. The fundamental economic agents in the television ecosystem are therefore the television companies ("broadcasters"), who organize daily, weekly, and annual schedules.

This demand and supply of television communication is only exceptionally regulated by the price. In fact, from an economic point of view, the television product is an "information good" (Shapiro and Varian 1998), which can be reproduced and distributed somewhat inexpensively. In the case of free-to-air television, be it state-owned or private, the good offered is a classic "public good." It is immaterial, it is not destroyed when consumed, and it is not divisible.

Its cost of production is not dependent on the number of people who will use it, that is, who will watch the program. However, the cost of production may affect the number of people who might want to look at the program (the most expensive productions will probably attract the larger audiences), but the cost will be the same, whether the program is seen or not. As "information goods," they are subject to economies of scale and scope, resulting from the high fixed costs of production of the first copy ("sunk costs"). The marginal cost of this product is equivalent to zero, since a viewer added to the program will not alter in any way the production costs to be incurred. Moreover, the fact that a person is watching the program does not diminish the ability of another person to see it. Since it is evident that no private operator would be interested in producing such a collective good as "television communication," unable to achieve earnings from the end user (the viewer), another market intersects with the demand and supply just described, permitting a return on investments: the advertising market.

Advertising

A second product generated by the television industry is, therefore, the "commercial," requested by companies that produce goods and services to be advertised. The viewer is characterized in this case as a receiver of advertisements, but also as a "product," offered by the broadcaster to its clients, such as the companies wishing to advertise on its TV channels. The price is the instrument that regulates this market, and it is influenced by variables that are related to the economic cycle, the market structure, and by regulation at national or supranational level (such as the European level). Different information about the viewers (age, gender, jobs, buying habits, and tastes), different kinds of advertising spaces (commercial breaks, sponsorship, product placement, etc.), and types of television programs (sport, news, current affairs, cartoon, sit-coms, TV serials, soaps, TV movies, reality, etc.) influence the value and price of each advertisement sold.

The television ecosystem is thus composed also by advertising agencies and companies that measure television ratings.

Within the EU, the AVMSD set a broad framework concerning some general rules applicable to all forms of commercial communications for audiovisual media services, both linear or non-linear, thus influencing the way broadcasters may collect their principal resource. First of all, all forms of audiovisual commercial communications for cigarettes and other tobacco products as well as audiovisual commercial communication for medicinal products and medical treatment available only on prescription are prohibited, and audiovisual commercial communications for alcoholic beverages shall not be aimed specifically at minors and shall not encourage

immoderate consumption. Then, more in particular, all audiovisual commercial communications shall be limited both in contents and in time; they shall be readily recognizable as such, distinguishable from editorial content (by optical and/or acoustic and/or spatial means) and subliminal techniques are prohibited all over the EU. The forms of advertising which are regulated include teleshopping, sponsorship, and product placement. Sponsorship includes any contribution made by public or private undertakings or natural persons not engaged in providing audiovisual media services or in the production of audiovisual works, to the financing of audiovisual media services or programs with a view to promoting their name, trademark, image, activities, or products. News and current affairs programs shall not be sponsored and also stricter rules may apply during documentaries and religious and children's programs. Product placement is a form of commercial where there is no payment but only the provision of certain goods or services free of charge, with a view to their inclusion in a program, and it is admissible in cinematographic works, films, and series made for audiovisual media services, sports, and light entertainment programs, with an exemption for children's programs.

Public Funded Models

However, advertising-based TV is not the only economic model in the television industry (Picard 2002). In some countries, and especially for state-owned public service broadcasters, a television license (or a broadcast receiving license) is required of the viewers for the reception of the TV service or the possession of a television set. This form of funding is typical of European countries (Josephine 2015), where Public Service Media (PSM) operate according to some national and supranational laws. For instance, in the UK, the British Broadcasting Corporation, established in 1922, has been broadcasting TV since 1936 and the 1954 Television Act established commercial television; in Italy, RAI, Radiotelevisione italiana, began a regular state-owned television service in 1954; in 1956, TVE, Televisión Española, started regular broadcasting in Spain. In 1967 the US Congress created the Public Broadcasting System, a noncommercial, public television network, funded, however, by congressional appropriations, viewer donations, and private corporate underwriters. During last years, many PSM objectives' revisions occurred, trying to adapt the infrastructure and the contents of historical PSM to multichannel and cross-platform scenarios. At the same time, all the mandates' revisions had to comply with the public service objectives identified in the "Protocol to the Amsterdam treaty on the system of public broadcasting in the Member States" (signed on November 10, 1997), which states that the system of public broadcasting (now PSM) in the Member States is directly related to the democratic, social, and cultural needs of each society and may help to preserve media pluralism. However, economic crisis as well as changing consumption habits both influenced the structure, the governance, and the business models of such services.

In Spain, for instance, advertising and public funded mixed models have been outmatched by public funds and private broadcasters/telecom operators taxation (Ley 7/2010). The same happened in France (this form of taxation has been approved by the Court of Justice of the European Union in 2013). In Finland, instead, direct taxes apply to the citizens, instead of a license, since 2012. In United Kingdom, finally, since April 2017, the independent authority OFCOM compassed the so-called BBC Trust in supervising the implementation of broadcasting legislation over the PSM.

As regards the public funded television business model, within the EU, article 107 of the Treaty on the Functioning of the European Union (TFEU) (ex Article 87 of Treaty establishing the European Community – TEC) provides that any aid granted by a Member State or through State resources in any form, which distorts or threatens to distort competition by favoring certain undertakings or the production of certain goods, shall be incompatible with the internal market (see entry ▶ "State Aids and Subsidies"). However, Article 106 specifies a number of circumstances (so-called "exemptions") in which the same state aid is acceptable, when some conditions apply, such as the service is clearly of general economic

interest by the EU county concerned, the undertaking in question must be explicitly entrusted by the EU country with the provision of that service, and the ban on state aid must obstruct the performance of the particular tasks assigned to the undertaking. Among these exemptions, some have direct or indirect influence on TV market. As for the direct influence, the public broadcasting service remit, in some circumstances and granting transparency in providing that service, may be included within the exemptions, according to the Amsterdam Protocol. In 2009 the EU Commission published a "Communication on the application of State aid rules to public service broadcasting," following a 2001 first Communication and then three rounds of public consultations (2008 and 2009), providing again for a derogation for funding granted to broadcasting organizations for the fulfillment of the public service remit so long as it does not affect trading conditions and competition in the EU to an extent that would be contrary to EU interests.

As for the indirect influence, the Commission adopted a revised General Block Exemption Regulation (GBER) in May 2014, which includes aid schemes for audiovisual works. Article 54 of the Commission Regulation lists the specific compatibility criteria applicable to aid schemes for audiovisual works: particularly, the aid must support the script-writing, development, production, distribution, and promotion of audiovisual works, assuming that those products are cultural products as defined by each Member State with a selection and a predetermined list of cultural criteria. Aid may take the form of aid to the production, preproduction, and distribution of audiovisual works. Same Regulation defines also (Article 2) what is considered a "difficult audiovisual work," which is "identified as such by Member States on the basis of pre-defined criteria when setting up schemes or granting the aid and may include films whose sole original version is in a language of a Member State with a limited territory, population or language area, short films, films by first-time and second-time directors, documentaries, or low budget or otherwise commercially difficult works."

Finally, another indirect influence of the EU state aid framework on the TV market is the exemption concerning public investments on broadband (see entry on ▶ "Public Investments: Broadband"), as the always more convergent audiovisual world is strongly shifting from terrestrial free-to-air television systems or satellite technology to a broadband-driven audiovisual consumption.

Subscription-Based Models

Finally, TV revenues may also come from channel/bundles/content subscriptions, paid by the audience to access the content, both linearly and on demand. This form of revenue accounts for more than a half of the total revenues of the TV market worldwide, in 2016 (Ofcom 2017). Pay-TV revenue is a well-established business model, especially in some country such as USA and Australia, followed by UK and Germany, while in other countries public funded or advertising-based models prevail.

Mostly, recently released films or TV series and sport events (so-called "premium contents") and back-catalogue films are the most popular types of programs watched on subscription services. These services are in charge of traditional broadcasters, who differentiate business models reorganizing the offer through different platforms and revenue models, or are provided also by Over-the-Top operators. The widening of the TV market encompasses unquestionably the enrichment of the pay-TV offer, and the competition between actors in search for the richest targets (who may pay to access content) is growing, also thanks to the convergence between technologies and platforms.

Future Directions

During the last years, the TV market has been changing completely and viewers can now access TV content directly, on Internet platforms. At the same time, more traditional business models – advertising based or public funded – persist, offering to the audience an incredible amount of audiovisual contents, linear and on demand. Despite the convergent television-Internet-

telecommunications world that has been expected for decades, integration between platforms was initially very difficult. Relations between the TV ecosystem actors have been rough. Telecoms operators (see entry on ▶ "Telecommunications") have tried to enter the television business many times, but it was not always easy and the first versions of Internet protocol television (IPTV) did not obtain their estimated success, in many countries. However, nowadays the so-called Quadruple players provide content across mobile, video, and broadband platforms (Ofcom 2017). Content producers, distributors, and broadcasters, finally, compete today both with telecoms operators and with players coming from the web, such as the so-called Over-the-Top players (Google, Amazon, Netflix, Hulu, . . .), offering to the audience new forms of services and a huge amount of audiovisual contents.

Cross-References

- ▶ Media
- ▶ Public Investments: Broadband
- ▶ State Aids and Subsidies
- ▶ Telecommunications

References

Adda J, Ottaviani M (2005) The transition to digital television. Econ Policy 20(41):160–209

AVMSD 2010. Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services

Charter of Fundamental Rights of the European Union (2012/C 326/02)

Commission communication on the application of state aid rules to public service broadcasting [Official Journal C 257 of 27.10.2009]

Commission Regulation N. 651/2014 of 17 June 2014 declaring certain categories of aid compatible with the internal market in application of Articles 107 and 108 of the Treaty

Communication from the Commission on the application of State aid rules to public service broadcasting [Official Journal C 257 of 27.10.2009]

Consolidated version of the Treaty on the Functioning of the European Union (2010/C 083/01)

Council Directive 89/552/EEC of 3 October 1989 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the pursuit of television broadcasting activities

European Audiovisual Observatory (Council of Europe) (2017) Audiovisual services in Europe – focus on services targeting other countries, Strasbourg

European Convention for the Protection of Human Rights and Fundamental Freedoms, amended (1950)

European Parliament resolution of 21 May 2013 on the EU Charter: standard settings for media freedom across the EU (2011/2246(INI))

Freedman D (2008) The politics of media policy. Polity, Cambridge, UK

Josephine MA-C (2015) Public service media remit in 40 European countries. IRIS bonus 2015-3. European Audiovisual Observatory, Strasbourg

Ley 7/2010, de 31 de marzo, General de la Comunicación. Audiovisual. Jefatura del Estado. "BOE" núm. 79, de 1 de abril de 2010

Ofcom (2016) The International Communications Market 2016

Ofcom (2017) The International Communications Market 2017

Owen BM, Wildman SS (1992) Video economics. Harvard University Press, Cambridge, MA. La Editorial, UPR

Picard R (2002) The economics and financing of media companies. No. 1. Fordham University Press, New York

Shapiro C, Varian HR (1998) Information rules: a strategic guide to the network economy. Harvard Business Press.

Terror

▶ Terrorism

Terrorism

Friedrich Schneider[1] and Daniel Meierrieks[2]
[1]Department of Economics, Johannes Kepler University of Linz, Linz-Auhof, Austria
[2]Department of Economics, University of Freiburg, Freiburg, Germany

Synonyms

Political violence; Terror

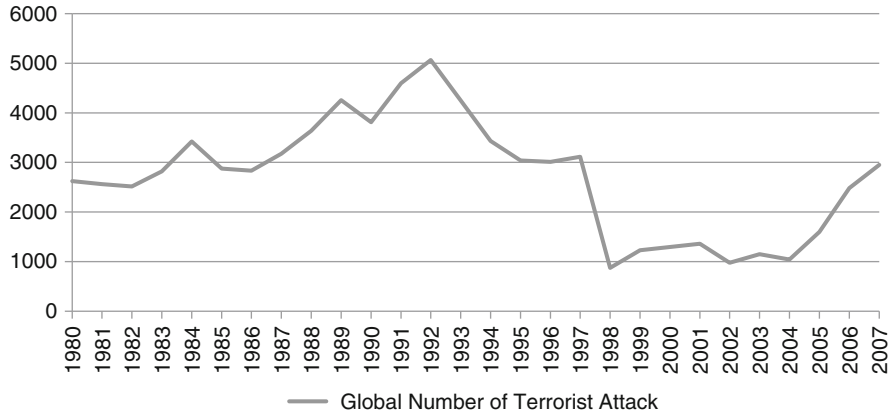
Definition

Terrorism is the premeditated use, or threat of use, of extra-normal violence by non-state outside the context of legitimate warfare with the intention to coerce and intimidate an audience larger than the immediate victims in order to obtain political, economic, religious, or other social objectives through intimidation or fear.

Introduction

The *Global Terrorism Database* (*GTD*) defines terrorism as any action by a non-state actor (usually, a terrorist organization) outside the context of legitimate warfare with the intention to communicate (through the use of violence) with, coerce, or intimidate an audience larger than the immediate victims of a terrorist act, where this act is associated with achieving political, economic, religious, or other social objectives. Especially the terrorist attacks on New York and Washington, D.C., on September 11, 2001, have sparked a renewed interest in the economic analysis of terrorism and counterterrorism.

Given the persistence of terrorism (illustrated by Fig. 1 which uses data from the *GTD* to illustrate the global patterns of terrorism over the past decades), we want to provide an overview of the academic literature on the *economics of terrorism and counterterrorism*. The following contribution is a condensed version of Schneider et al. (2014) and uses material from it. In the section "Rational-Choice Theory and the Economic Analysis of Terrorism" of this contribution, we discuss the role of rational-choice theory in the economic analysis of terrorism and counterterrorism. Here, we argue that simple cost-benefit models using rational-choice representations of terrorist behavior provide a well-founded model for the study of terrorism. In the section "Terrorism and Counterterrorism in a Rational-Choice Framework," we review the fundamental strategies of counterterrorism that follow from the rational-choice framework. We discuss the implications for the design of appropriate counterterrorism efforts and the empirical evidence assessing the effectiveness of these efforts. Here, we also take stock of the literature on the causes and consequences of terrorism. However, we also hint at the limitations associated with these simple cost-benefit models, which especially relate to the failure of accounting for collective action problems linked to the phenomenon of transnational terrorism and the dynamic interaction between terrorism and counterterrorism. Thus, in the section "Interaction Between Terrorism and Counterterrorism," we discuss some of the consequences that result from the reaction of terrorists and other economic agents to distinct counterterrorism measures. The section "Conclusion and Future Research" concludes by discussing which counterterrorism strategies



Terrorism, Fig. 1 Global terrorist activity, 1980–2007

may ultimately prove most helpful in the fight against terrorism.

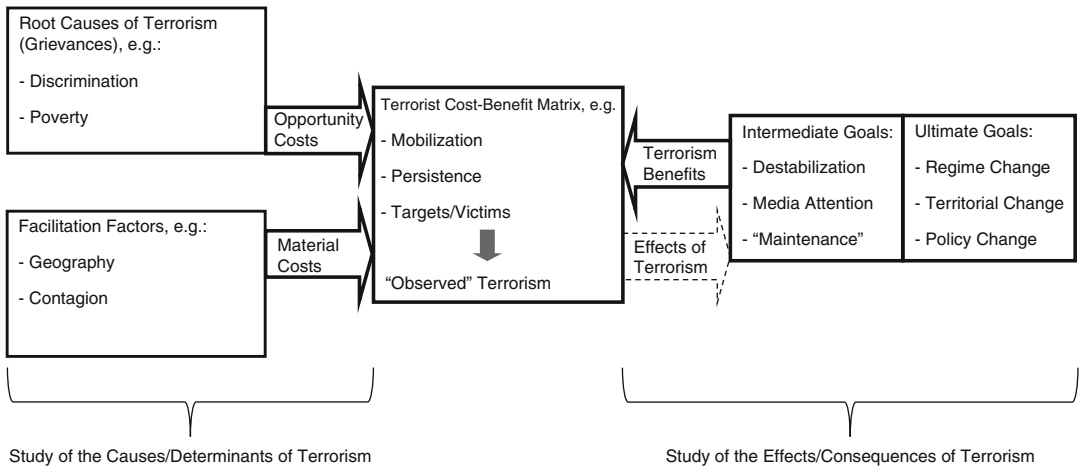
Rational-Choice Theory and the Economic Analysis of Terrorism

Rational-choice models are the theoretical workhorse of most economic analyses of terrorism (e.g., Landes 1978; Sandler and Enders 2004; Caplan 2006; Freytag et al. 2011). In short, in a rational-choice approach, terrorists are considered rational actors who choose the optimal (utility-maximizing) level of violence by considering the costs, benefits, and opportunity costs of terrorism, where the utility from terrorism is usually associated with achieving certain (political) goals (Sandler and Enders 2004).

The existence of a calculus of (rational) terrorists allows for an economic analysis of terrorism. For one, it informs empirical studies on the *determinants of terrorism*. For another, it informs empirical studies on the *consequences of terrorism*. Figure 2 shows the relationship between rational-choice theory and the study of terrorism and counterterrorism.

Terrorists as Rational Actors

As argued by Krueger and Maleckova (2003), rational-choice models of terrorism can be considered as an extension of models that economically analyze criminal behavior. Economic models of crime (Becker 1968) suggest that criminals are rational actors who maximize their utility subject to a calculus that involves the costs of criminal activity (e.g., from the risk of punishment), its benefits (e.g., the "wage" a crime pays), and its opportunity costs (e.g., foregone earnings from noncriminal activity). Analogous to this, rational-choice models of terrorism assume terrorists to be rational actors – behaving more or less like *homi economici* – that try to maximize their utility, given the benefits, (opportunity) costs, and economic constraints linked to these actions. While public perception tends to view terrorist behavior as "irrational," psychological studies of terrorist behavior provide little evidence that terrorists routinely suffer from mental incapacities (for a review, see Victoroff (2005)). As a matter of fact, Caplan (2006) provides an extensive analysis of terrorist irrationality and comes to the conclusion that the sympathizers of terrorism and most active terrorists act more or less rational, so that "the rational choice model of terrorism is not that far from the truth [and] the Beckerian analysis of crime remains useful" (Caplan 2006, p. 101). This may even extent to the (rare) case of *suicide terrorism*. While Caplan (2006) argues that suicide terrorists indeed typically violate the rationality assumption of the rational-choice model and thus should be considered as outliers, Pape (2003) argues that even



Terrorism, Fig. 2 Rational-choice theory and the study of terrorism

suicide terrorism is the result of strategic-rational decision-making.

The Terrorists' Calculus

The Direct Costs Associated with Perpetrating Terrorism

The direct or material costs of terrorism are one element of the terrorists' calculus. Usually, these costs are associated with the operations of a terrorist organization, i.e., offensive and defensive activities. For instance, they accrue from acquiring financial resources, purchasing firearms and explosives, and establishing safe houses to evade government punishment.

The Benefits of Terrorism

Benefits of terrorism are connected with the tactical and strategic goals of terrorism. As argued by Schelling (1991), the short-run (tactical) goals of terrorism include *politico-economic destabilization* (to weaken their enemy, i.e., the state) and *media attention* (to communicate the terrorists' cause). Thus, achieving these tactical goals ought to provide the member of a terrorist group with benefits. The long-run goals of terrorist groups usually involve the wish to induce political, economic, religious, or social change, most prominently (for separatist terrorist groups) gaining independence and (for religious and left-wing or right-wing terrorist groups) changing a country's politico-economic system to have it coincide with religious doctrine or political ideology. These political objectives are discussed in more detail in Shughart (2006). Any success related to achieving these long-run goals (through government concessions, territorial gains, winning political influence, etc.) also ought to constitute a benefit from terrorist activity.

The Opportunity Costs of Terrorism

The opportunity costs of terrorism refer to the foregone utility associated with non-terrorist activity. Typically, this utility comes from material rewards (i.e., wages) linked to nonviolence, e.g., participation in the ordinary economic life. It can also be understood as a monetary equivalent (derived from economic activity) of the potential political influence associated with terrorist activity. Such lines of reasoning are at the center of many contributions that try to economically model terrorist behavior (e.g., Sandler and Enders 2004; Freytag et al. 2011).

Terrorism and Counterterrorism in a Rational-Choice Framework

The identification of an economic calculus associated with terrorist behavior allows for *three fundamental strategies* to reduce terrorist activity. Ceteris paribus, the utility-maximizing level of terrorism (i.e., the observed level of violence) ought to be lower when (1) the costs of terrorism are increased, (2) the benefits of terrorism are reduced, or (3) the opportunity costs of terrorism are raised. We discuss these options below in more detail.

Raising the Costs of Terrorism

Any counterterrorism policy that aims at raising the material costs of terrorism (e.g., by raising the penalty for terrorist offenses, direct police or military efforts) ought to make it more difficult for terrorist groups to maintain their level of activity.

The prevailing counterterrorism strategy to raise the direct costs of terrorism is *punishment and deterrence*. This strategy involves the use of direct state action by the police, military forces, and intelligence agencies to capture active terrorists and their supporters, while also deterring potential recruits (due to long prison sentences, increased probability of being captured, etc.). Indeed, counterterrorism activities by the police and intelligence services (e.g., infiltration through the use of informers and undercover agents, observation, information gathering) have repeatedly weakened the operative capacity of terrorist organizations (for an overview, see Schneider et al. (2014)). Also, empirical evidence suggests that increased punishment for specific kinds of terrorist activity leads to fewer of these acts (e.g., for the case of skyjackings, see Landes (1978)). As for a more harsh direct counterterrorism strategy, efforts in the form of decapitation – i.e., the killing of terrorist leaders – have also been shown

to prove helpful against terrorist groups (e.g., Johnston 2012).

Reducing the Benefits of Terrorism

Previously, we have established that the benefits of terrorism are directly linked to the short-run and long-run objectives of terrorist organizations. In the short run, the benefits from terrorism arise from politico-economic destabilization and media attention, both of which help the terrorists to achieve their long-run sociopolitical objectives. Indeed, many empirical studies have established that terrorism tends to negatively affect political development (e.g., in the form of respect for human rights) and political stability (e.g., Dreher et al. 2010). There is also evidence that terrorism reduces economic activity, e.g., by depressing domestic and foreign investment or tourism (e.g., Crain and Crain 2006; for a review, see Sandler and Enders (2008)).

Counterterrorism efforts may be effective when they successfully deny terrorist groups media attention and increase a country's politico-economic resiliency to terrorism's destabilizing effects, which ultimately ought to mean that political concessions associated with the long-run goals of terrorist groups become less likely. This is because less media attention and less success in producing politico-economic damage are expected to negatively affect a terrorist organization's bargaining position vis-à-vis the government it opposes (Schneider et al. 2014).

One interesting strategy to increase politico-economic robustness is *decentralization*. Political decentralization ought to make it less likely that terrorism creates a political vacuum (e.g., when prominent political figures are assassinated) that cannot be filled by other actors or levels of government (Frey and Luechinger 2004). Economic decentralization is expected to have a similar effect on terrorism. It may include avoiding concentrating power within a company in the hands of few individuals (because these individuals are attractive targets for attacks), not concentrating a company in one large headquarters (particularly when this headquarter is located in an iconic building such as the World Trade Center), and using multiple (rather than monopolistic) suppliers to immunize a company's supply chain against disruptions (Frey and Luechinger 2004). Indeed, Dreher and Fischer (2010) find that decentralization may reduce the likelihood of terrorism.

Other means to reduce the benefits of terrorism may include, e.g., the increased protection and fortification of prospective terrorist targets and denying the terrorists media attention by means of spreading disinformation (Schneider et al. 2014).

Influencing the Opportunity Costs of Terrorism

While raising the direct/material costs of terrorism usually involves manipulating related *facilitation factors* (e.g., by restricting access to firearms and explosives), influencing the opportunity costs of terrorism typically involves policies that try to ameliorate *grievances* that underlie social conflict.

Among the factors possibly involving grievances and thereby inciting terrorism that are analyzed in cross-national studies are (1) *poor socioeconomic conditions* (e.g., Freytag et al. 2011); (2) *unfavorable political and economic institutions* which, e.g., do not sufficiently protect civil liberties and do not enable politico-economic participation (e.g., Krueger and Maleckova 2003; Abadie 2006); (3) *demographic stress*, e.g., in the form of discrimination along ethno-religious lines (e.g., Piazza 2011); (4) trends in *globalization* that may generate resentment among the "globalization losers" and traditionalist who fear that the inflow of foreign ideas threatens their local culture or religion (Zimmermann 2011); and (5) aggressive *foreign policy behavior* – in the form of, e.g., military interventions, state sponsorship of terrorism, and overall politico-military dominance – which tends to coincide with more terrorist activity directed against the proponents of such policies (e.g., Pape 2003; Dreher and Gassebner 2008). Gassebner and Luechinger (2011) and Krieger and Meierrieks (2011) summarize the empirical literature on the causes of terrorism.

Ameliorating the aforementioned social conditions ought to reduce terrorism by swaying its opportunity costs in ways that make violence

a less attractive option. However, according to the reviews by Gassebner and Luechinger (2011) and Krieger and Meierrieks (2011), there is little consensus on the importance of specific social conditions in the emergence of terrorism. For instance, while some cross-country studies find that poor socioeconomic conditions predict terrorism (e.g., Freytag et al. 2011), other studies find that political variables are more important than economic ones (e.g., Krueger and Maleckova 2003; Abadie 2006). Thus, there seems to be no obvious "policy panacea" to fight terrorism by favorably affecting its opportunity costs. What is more, influencing the social conditions that underlie terrorist conflict in favorable ways tends to be a long-run and complex counterterrorism option, particularly in comparison to options that try to affect the direct costs of terrorism.

The Case of Transnational Terrorism

Counterterrorism efforts may be further complicated when terrorism becomes transnational. By definition, this is the case when more than one country is involved in a terrorist conflict. For instance, terrorism may internationalize when terrorist groups use foreign territory to challenge another government (e.g., the use of Syria, Jordan, Lebanon, Tunisia, and other Middle Eastern countries as bases of operation for Palestinian terrorist activity against Israel in the 1970s and 1980s).

International counterterrorism strategies still aim at influencing the (opportunity) costs and benefits of terrorism in ways that make (internationalized) terrorism less attractive. However, it is precisely their transnational nature that makes these strategies even more difficult to implement. Sandler (2005) shows that internationalized terrorism usually involves *collective action failures* for countries attacked by transnational terrorist groups. While governments are able to choose an optimal counterterrorism strategy in the fight against domestic terrorism, counterterrorism against transnational terrorism may involve externalities. For instance, it may be attractive for some – often weak – states to tolerate the activities of terrorist organizations within their borders in exchange for no direct harm at the expense of other nations (*paid riding*). Also, national governments may be tempted to overstress defensive counterterrorism measures to deter terrorist attacks, which may merely result in a relocation of terrorist attacks against the respective nation's citizens (e.g., there were more anti-US attacks in the Middle East and Africa after 9/11 had led to stronger counterterrorism efforts within the United States). Alternatively, as suggested by Sandler (2005), countries may try to place the burden of offensive counterterrorism measures (e.g., preemptive strikes) onto the prime target of transnational terrorism, thereby benefitting from the reduction of terrorism without paying for it (*free riding*). More generally, defensive policies can be regarded as largely private goods, so that the benefits of security provision are mostly internalized by the investors, while proactive policies show the characteristics of public goods (Sandler and Siqueira 2006). Consequently, this may lead to an oversupply of defensive and an undersupply of proactive measures (Sandler and Siqueira 2006).

Collective action problems (e.g., free and paid riding, benefits of noncompliance) have so far undermined many measures directed against terrorism at an international level. For instance, Cauley and Im (1988) find that the introduction of a United Nation's convention on preventing crime against diplomatic personnel did not help to reduce terrorism directed against diplomats. Similar problems arise when international counterterrorism efforts aim at curtailing terrorism by means of military actions (e.g., the US air strikes against Libya in 1988 in retaliation of alleged Libyan support for anti-US terrorism, which in turn led Libya to sponsor anti-American terrorism) and the use of economic and/or military aid (Schneider et al. 2014).

Interaction Between Terrorism and Counterterrorism

Previously, we have argued that rational-choice theory helps to understand how terrorism develops and how it can be countered through three fundamental strategies that aim at influencing the (opportunity) costs and benefits of terrorism in ways that make terrorism less attractive. One

shortcoming of this approach, however, is its *static* nature. In reality, terrorism and counterterrorism tend to interact. In this section, we therefore give a brief – nontechnical – summary of typical interactions between terrorism and counterterrorism. Some of these points are also discussed in Kydd and Walter (2002) and Findley and Young (2012).

Adaption, Innovation, and Substitution

Terrorist organizations may be able to adapt to specific counterterrorism measures by means of innovation and/or substitution. The former may include, e.g., the use of more advanced terrorist "technology" to attack (e.g., the use of more powerful weapons) or organizational innovation as a hierarchical group develops into a looser terrorist network with a cell structure. The latter may involve, e.g., choosing new targets that are less protected by counterterrorism measures. Arguably, substitution effects are standard elements of economic models, so that it is not surprising that such effects also matter to the economic study of terrorism and counterterrorism (Enders and Sandler 2004). As a consequence, the evidence suggests that some counterterrorism measures may merely change the face of terrorism (e.g., new attack methods, new targets) but not affect the overall level of terrorist violence (Schneider et al. 2014).

Provocation and Escalation

Findley and Young (2012) argue that terrorist groups may use terrorism to provoke a harsh, disproportionate response by the government (e.g., by means of excessive police force), which in turn is expected to fuel radicalization and social polarization, meaning that provocation and excessive counterterrorism measures by the state may easily result in cycles of violence. Indeed, it seems to be in the natural interest of terrorist groups to muster additional support by letting any social conflict escalate, given that broader conflicts (e.g., civil wars) make it easier for them to pursue and implement their agendas (Findley and Young 2012).

Spoiling the Peace

As a specific reaction to government concessions, terrorist groups may try to spoil the peace. For instance, terrorist groups may have economic incentives (e.g., gains from illegal activity) associated with conflict and thereby oppose peace. Thus, it may be in the natural interest of extremist factions to sabotage the peace (Kydd and Walter 2002). Crucially, such sabotage may result in – as it is in the interest of the attacking groups – an end of negotiations and concessions and, instead, provoke more violent counterterrorism measures. Kydd and Walter (2002) argue that spoiling the peace – as a reaction to benevolent policy measures such as concessions – may be particularly effective when mistrust and weakness among the more moderate negotiator on the sides of the terrorists and the state abound.

Organizational Evolution

We have already discussed above that terrorist groups may innovate in the face of counterterrorism by changing their internal organization (e.g., moving from a hierarchical to a cell structure). However, counterterrorism may also result in changing the overall politico-military orientation of a terrorist group. The nature of these changes depends on the offered incentives and the effectiveness of counterterrorism. For one, positive incentives and relative counterterrorism effectiveness may lead to a situation where a group gives up its armed struggle, evolving into a political party that tries to foster change nonviolently. One example is the Colombian *Movimiento 19 de Abril*. For another, however, negative incentives and/or relative counterterrorism ineffectiveness may lead to an escalation of conflict, so that terrorist campaigns may evolve into full-scale insurgencies (Findley and Young 2012). As another possibility, perhaps in frustration over the inability to achieve certain political goals due to counterterrorism, terrorist groups may become increasingly interested in purely financial gains, evolving into criminal groups.

Conclusion and Future Research

In this entry, we provided an overview of theoretical and empirical studies on the economic

analysis of terrorism and counterterrorism. We argued that rational-choice models of terrorist behavior provide a good starting point for such analyses. Here, simple cost-benefit models that follow from rational-choice approaches imply that terrorism can be fought by affecting the terrorists' calculus which involves the (opportunity) costs and benefits of terrorism. We discussed a number of specific counterterrorism strategies (e.g., direct police or military action, decentralization) that follow from such models. One factor that complicates these models – and the strategies derived from them – are collective action problems that arise when terrorism becomes transnational (the most infamous example being the 9/11 attacks). Another complicating factor associated with these simple cost-benefit models is that they are usually static and thus miss the dynamic nature of the terrorism-counterterrorism relationship. We (non-exhaustively) discussed a number of reactions of terrorist groups to specific counterterrorism strategies (e.g., the use of innovative modes of attack in the light of target-hardening efforts). These interactions show that the relationship between terrorism and counterterrorism tends to be more complex than simple cost-benefit models – even though they provide a good starting point for any economic analysis of terrorism – can capture.

Considering the most effective counterterrorism strategy, our entry suggests the following: (1) Strategies that involve influencing the benefits of terrorism (e.g., target protection) usually do not stop terrorism but lead to adaption (new targets, new methods, etc.). (2) Raising the direct/material costs of active terrorists (e.g., through military means) may prove effective in the short run. Indeed, many terrorist groups have been negatively affected by military pressure. However, such efforts may easily backfire, creating (potentially large) unintended consequences. Here, (relative) counterterrorism ineffectiveness may contribute to the emergence of powerful insurgencies or crime networks, i.e., to the evolution rather than decline of terrorist groups. For instance, the French counterterrorism efforts during the Battle of Algiers of 1957 against the Algerian *Front de Libération Nationale* (*FLN*) were a military success but – due to use of torture and other harsh counterterrorism measures – led to the international political isolation of France and growing (domestic and international) support for the *FLN*, while also triggering a political crises in France. (3) Strategies that aim at raising the (opportunity) costs of terrorism by "winning the hearts and minds" of would-be terrorists, potential terrorism supporters, and possibly even active terrorists seem to be more effective in the long run. For instance, terrorist groups may disappear due to a loss of legitimacy – e.g., as popular support for terrorism dwindles due to higher terrorism opportunity costs in the face of benevolent counterterrorism strategies – or evolve into political parties when political participation opportunities open up.

Clearly, the effectiveness of these strategies is very much context-dependent. Counterterrorism strategies that worked well against terrorism in the twentieth century (which usually had domestic goals, was hierarchically structured, and attacked primarily military targets) may prove unsuccessful against the *Al-Qaeda*-styled terrorism of the twenty-first century (which has transnational goals, is organized as a network, and also attacks civilian targets). This provides many avenues for future research. For instance, this research may try to evaluate the appropriateness of counterterrorism strategies vis-à-vis, e.g., the goals of terrorist organizations (to assess which kinds of concessions can be made), the implementation costs of specific policies (to examine their cost-efficiency), the international perspective (to factor in collective action problems), the overall level of popular support for terrorism, and the organizational structure and politico-military strategy of a terrorist group. These factors have been largely disregarded in the theoretical and empirical literature on terrorism and counterterrorism.

Cross-References

► Crime: Organized Crime and the Law
► Violence, Conflict-Related

References

Abadie A (2006) Poverty, political freedom, and the roots of terrorism. Am Econ Rev 96:50–56

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

Caplan B (2006) Terrorism: the relevance of the rational choice model. Public Choice 128:91–107

Cauley J, Im EI (1988) Intervention policy analysis of skyjackings and other terrorist incidents. Am Econ Rev 78:27–31

Crain NV, Crain WM (2006) Terrorized economies. Public Choice 128:317–349

Dreher A, Fischer JAV (2010) Decentralization as a disincentive for transnational terror? An empirical test. Int Econ Rev 51:981–1002

Dreher A, Gassebner M (2008) Does political proximity to the U.S. cause terror? Econ Lett 99:27–29

Dreher A, Gassebner M, Siemers LH-R (2010) Does terror threaten human rights? Evidence from panel data. J Law Econ 53:65–93

Enders W, Sandler T (2004) What Do We Know About the Substitution Effect in Transnational Terrorism? In Silke A (ed) Research on Terrorism: Trends, Achievements, and Failures. London, Frank Cass, pp. 119–137

Findley MG, Young JK (2012) Terrorism and civil war: a spatial and temporal approach to a conceptual problem. Perspect Polit 10:285–305

Frey BS, Luechinger S (2004) Decentralization as a disincentive for terror. Eur J Polit Econ 20:509–515

Freytag A, Krüger JJ, Meierrieks D, Schneider F (2011) The origins of terrorism: cross-country estimates of socio-economic determinants of terrorism. Eur J Polit Econ 27:5–16

Gassebner M, Luechinger S (2011) Lock, stock, and barrel: a comprehensive assessment of the determinants of terror. Public Choice 149:235–261

Johnston PB (2012) Does decapitation work? Assessing the effectiveness of leadership targeting in counterinsurgency campaigns. Int Secur 36:47–79

Krieger T, Meierrieks D (2011) What causes terrorism? Public Choice 147:3–27

Krueger AB, Maleckova J (2003) Education, poverty and terrorism: is there a causal connection? J Econ Perspect 17:119–144

Kydd A, Walter BF (2002) Sabotaging the peace: the politics of extremist violence. Int Organ 56:263–296

Landes WM (1978) An economic study of US aircraft skyjackings, 1961–1976. J Law Econ 21:1–31

Pape RA (2003) The strategic logic of suicide terrorism. Am Polit Sci Rev 97:341–361

Piazza JA (2011) Poverty, minority economic discrimination, and domestic terrorism. J Peace Res 48:339–353

Sandler T (2005) Collective versus unilateral responses to terrorism. Public Choice 124:75–93

Sandler T, Enders W (2004) An economic perspective on transnational terrorism. Eur J Polit Econ 20:301–316

Sandler T, Enders W (2008) Economic consequences of terrorism in developed and developing countries: an overview. In: Keefer P, Loayza N (eds) Terrorism, economic development, and political openness. Cambridge University Press, Cambridge, pp 17–47

Sandler T, Siqueira K (2006) Global terrorism: deterrence versus pre-emption. Can J Econ 39:1370–1387

Schelling TC (1991) What purposes can "international terrorism" serve? In: Frey RG, Morris CC (eds) Violence, terrorism, and justice. Cambridge University Press, Cambridge, pp 18–32

Schneider F, Brück T, Meierrieks D (2014) The economics of counter-terrorism: a survey. J Econ Surv (forthcoming), https://doi.org/10.1111/joes.12060

Shughart WF (2006) An analytical history of terrorism, 1945–2000. Public Choice 128:7–39

Victoroff J (2005) The mind of the terrorist: a review and critique of psychological approaches. J Confl Resolut 49:3–42

Zimmermann E (2011) Globalization and terrorism. Eur J Polit Econ 27:152–161

Further Reading

Cronin AW (2011) How terrorism ends: understanding the decline and demise of terrorist campaigns. Princeton University Press, Princeton

Enders W, Sandler T (2011) The political economy of terrorism. Cambridge University Press, New York

Hoffman B (2006) Inside terrorism. Columbia University Press, New York

Walker C (2011) Terrorism and the law. Oxford University Press, Oxford

Third-Party Financed Litigation

► Third-Party Litigation Funding

Third-Party Financing of Litigation

► Third-Party Litigation Funding

Third-Party Litigation Funding

Myriam Doriat-Duban
Department of Economics, BETA UMR 7522, Université de Lorraine, Nancy, France

Synonyms

Third-party financed litigation; Third-party financing of litigation

Abstract

Third-party litigation funding allows the transfer by the plaintiff of his/her legal costs to a financial company whose sole aim is to make profit. TPLF have an impact on access to justice but also on conflict resolution and finally on social well-being. The aim is to show how law and economics scholars invest this new field of the conflict literature and study the role of this new actor in the litigation.

Definition

Third-party litigation funding (TPLF) is a financial activity that consists in an investment company specialized in the financing of litigation agreeing to cover all or part of the trial costs of a litigant, in exchange for a portion of the damages paid if the case is won. The remuneration of the funding third party can be a multiple of the initial investment (1.5–6 times), a percentage of the award obtained in a judgment or out-of-court settlement (20–40%, sometimes 50%), or a combination of both (Veljanovski 2012). The funds used are provided by wealthy individuals, families or institutions, insurance companies, or hedge funds (de Silguy 2013; Veljanovski 2012). The return/risk ratios can reach very high levels, up to 200% of the amounts invested (McLaughlin 2007), sometimes more (Abrams and Chen 2013). There are several types of TPLF: interest-bearing loans intended to provide replacement income to victims of bodily injuries or discriminatory practices (consumer legal funding), loans to firms of lawyers in a contingent fees system, and investment in commercial litigation in exchange for a percentage of the damages (Faure and De Mot 2012).

Introduction

Historically, TPLF first appeared in the Middle Ages, but the prohibition of "maintenance" and "champerty" in the United States prevented it developing (Lyon 2010). It reappeared in Australia at the beginning of the 1990s, before spreading to other common law countries. Its development in countries with a civil law tradition came later and was more modest, with the exception of Germany and, to a lesser extent, Austria and Switzerland. TPLF concerns both litigation before the courts and arbitration procedures (in international trade in particular).

There are many arguments in favor of developing TPLF. The most immediate advantage is the removal of the financial barrier to the plaintiff's access to justice (Jackson 2009). Other arguments relate to the easier access to third-party capital (Daughety and Reinganum 2014); better allocation of funds for companies, which can invest in their business instead of devoting funds to their defense (Rubin 2011; Veljanovski 2012); and a better division of the risks (De Morpurgo 2011) between a risk-averse plaintiff and a neutral third-party funder who can pool the risks in a portfolio of claims not correlated with each other (Rubin 2011).

The economic analysis of TPLF means studying the impact of having litigation funded by a third party on the one hand, litigants' access to justice on the other, the method of resolving disputes, and finally the impact on social well-being.

TPLF and Access to Justice

From a theoretical point of view, TPLF is based on the idea that the right to sue is exchangeable according to the Coase theorem: by injecting funds into the case, the funder purchases the right to receive all or part of the recovery. By doing so, the funder facilitates access to justice by removing the financial constraint on the plaintiff, as do other schemes (legal aid, legal expenses insurance, contingent fees), but only in lawsuits where the stakes are high (De Fontmichel 2012) and mainly, in Europe, in the field of international arbitration. In a model inspired by Katz (1990), Deffains and Desrieux (2014) have demonstrated, however, that access to justice is not guaranteed even for meritorious claims. Indeed, the third-party funder will invest in a case if, first of all, he considers that it is sufficiently profitable once all the transaction costs have been taken into account (obligation to sign two contracts, one between the lawyer and the client and another between the third part funder and the client, cost

of negotiating the percentage of the damages or the interest rate, costs of obtaining experts' opinions, dispute solving costs, etc.) and, secondly, if it is more profitable than alternative investments. They also show that where there is asymmetry in the information on the real quality of the case, TPLF can encourage plaintiffs bringing frivolous lawsuits to bring more actions than in a no-win, no-fee arrangement or if they are self-financing. This outcome is due to the fact that the equilibrium settlement amount is lower with TPLF and therefore the likelihood of reaching an arrangement with the defendant is higher.

As well as facilitating access to justice, TPLF guarantees a better defense of plaintiffs' interests. Indeed, TPLF will orient the plaintiff toward better lawyers, whom he will be also be able to better control (Schanzenbach and Dana 2009). However, a problem of agency can arise between the third-party funder and the lawyer. Demougin and Maultzsch (2014) nevertheless show that it is possible to arrive at a combination of third-party funding and a no-win, no-fee arrangement that can overcome both these agency issues and the access to justice difficulties of potential plaintiffs with "deserving" claims.

Impact of TPLF on Methods of Conflict Resolution

TPLF has an impact on the method of conflict resolution because it increases the plaintiff's bargaining power: with the benefit of substantial financial resources thanks to the support of the third-party funder, his threat of taking legal action gains in credibility (De Morpurgo 2011). TPLF is therefore thought to be a way of enabling small- and medium-sized enterprises to fight large international disputes on an equal footing (De Fontmichel 2012).

Furthermore, TPLF breaks the monopsony power of the defendant: the third-party funder, along with the defendant who wants to come to an arrangement, can be considered as buyers of the plaintiff's right to sue (Hylton 2014). With the resources that he devotes to raising the value of the right he has thus acquired, the third-party funder enables the plaintiff to obtain a better award (Avraham and Wickelgren 2014) and ensures a higher likelihood of judgments being enforced (De Fontmichel 2012).

The credibility of the plaintiff's threat of legal action is also reinforced by the credibility of the signal concerning the quality of the case, due to the in-depth investigations (defendant's solvency, chances of success, quality of the lawyers) carried out by the third-party funder in order to minimize his risks (Faure and De Mot 2012). For Avraham and Wickelgren (2014), it is in the third-party funder's interests to send a clear signal to the defendant and to the judge about the quality of the case he is funding, via the interest rate applied to the plaintiff.

Ultimately, few claims will be selected (those with more than a 70% likelihood of being won, according to Veljanovski 2012), so that the third-party funder's actual risk is zero. Daughety and Reinganum (2014) confirm this in the case of TPLF where the third party provides the plaintiff with funds to cover his day-to-day expenses. The authors have produced a signal model where the plaintiff has private information on the real value of his case that determines its type. They show that the optimal interest rate, understood as that which maximizes the joint payment of the plaintiff and the funder, always leads to an out-of-court settlement of the dispute, whatever the type of plaintiff. Kirstein and Rickman (2004) arrive at the same result in a configuration where there is no information asymmetry but only diverging estimates by the parties of the plaintiff's chances of winning the court case.

TPLF and the Improvement of Social Well-Being

By facilitating access to justice, TPLF generates positive externalities that tend to increase social well-being. Thus, Hylton (2012) explains that if socially desirable legal action is not taken because it is not cost-effective, then TPLF is socially desirable because the transfer of costs that it allows makes legal action cost-effective for the plaintiff from a private point of view. This implies that there are an optimum number of litigation cases. Any analysis of the social desirability of TPLF should therefore start by determining whether the number of litigation

proceedings brought is lower or higher than the socially optimal number.

Other positive external effects are also discussed in the literature. For example, TPLF is said to increase competition between lawyers (Veljanovski 2012; Costargent 2012) but also between large law firms on the one hand and third-party funders on the other (De Fontmichel 2012). Finally, according to De Morpurgo (2011), by facilitating access to justice, TPLF is said to increase equality of litigants before the law, which would therefore have the consequence of increasing the usefulness of all individuals with a concern for justice and therefore also social well-being.

In addition to these positive external effects, there are also negative external effects, which render the conclusions on the social desirability of TPLF theoretically undetermined. It is quite possible, for example, that the drop in the number of expected damages cases will not last. Hylton (2014) is interested not in existing, matured claims but in potential future, unmatured claims. The sale of ex ante rights is a complex issue because inefficient rights transfers can lead to an increase in the number of damages cases for potential victims. It all depends on the third party's intentions. If what he wants to achieve is the payment of damages, his interests are in line with those of the potential victim, and social well-being is therefore potentially increased. However, if his aim is to avoid proceedings by purchasing his right to reparation from the victim (interests in line with those of the defendant), then there will be fewer precautions taken and therefore an increase in the frequency of damages cases.

Rubin (2011), however, worries about the defendant's situation. The defendant's participation in the conflict is determined by the plaintiff's decision to take legal action. The American rule on the allocation of litigation costs creates an externality for the defendant, who necessarily has to bear his litigation costs. In the case of TPLF, in particular in commercial lawsuits, those litigation costs are usually high. Any measure facilitating plaintiffs' access to justice must take account of the costs borne by the opposing party. In the end, better access to justice is socially desirable if the expected benefits are higher than all the litigation costs, as for it the damages paid are a simple transfer.

Conclusion

The question of the socially desirable nature of TPLF runs through all the articles on this form of funding access to justice. The fears most often expressed concern the risk of encouraging socially undesirable litigation; economic analysis of conflicts shows that this appears to be unfounded due to third-party funders' very strict selection of the cases they take on. Furthermore, there can be other positive effects, in particular in terms of damage prevention. However, these are counterbalanced by the risk of congestion of the courts, the development of a compensation culture rather than a culture of recognizing rights and the risks that the intervention of a third party can bring to bear on the lawyer-client relationship.

It has to be admitted that the theoretical conclusions do not provide a definitive answer to the question of the social desirability of this type of funding. Empirical studies could in this case provide some assistance with decision-making. There is only one empirical article on the subject, concerning Australia. Abrams and Chen (2013) demonstrate that third-party funding of litigation increases the frequency of legal action being brought and the number of court cases. They ascertain in particular that in the Australian States where third-party litigation funding is common, the courts are very congested, the number of cases is completed, and the elucidation rate are lower, which also has consequences for the courts' spending. And yet, the authors explain that these effects may be temporary and the conclusions in terms of social well-being ambiguous (increase in the rate of out-of-court settlements, faster recognition of rights, clarification of the law). These conclusions cannot, however, be generalized to all countries.

Cross-References

▶ Legal Aid
▶ Litigation and Legal Expenses Insurance

References

Abrams DS, Chen DL (2013) A market for justice: a first empirical look at third party litigation funding. Univ PA J Bus 14(7):1075–1109. https://doi.org/10.2139/ssrn.2404483

Avraham R, Wickelgren A (2014) Third-party litigation funding – a signaling model. De Paul Law Rev 63:233–263. https://doi.org/10.2139/ssrn.2302801

Costargent J-R (2012) Le financement par un tiers comme réponse aux évolutions de l'arbitrage international, avec Guy Lepage, Versailles International Arbitration and Business Law Review. Disponible à l'adresse: http://www.lafrancaise-icfund.lu/fileadmin/redacteurs/docs/articles/article_financement_par_tiers_J-C_Costargent.pdf

Daughety AF, Reinganum JR (2014) The effect of third-party funding of plaintiffs on settlement. Am Econ Rev 104(8):2552–2566. https://doi.org/10.2139/ssrn.2197526

Deffains B, Desrieux C (2014) To litigate or not to litigate ? The impacts of third-party financing on litigation. Int Rev Law Econ. https://doi.org/10.1016/j.irle.2014.08.005. Available online: https://doi.org/10.1016/j.irle.2014.08.005

De Fontmichel M (2012) Les sociétés de financement de procès dans le paysage juridique français. Revue des sociétés 8:279

De Morpurgo M (2011) A comparative legal and economic approach to third-party litigation funding. Cardozo J IntL Comp Law 19:343–412

Demougin D, Maultzsch F (2014) Third-party financing of litigation: legal approaches and a formal model. CESifo Econ Stud 60:525–553. https://doi.org/10.1093/cesifo/ift006

De Silguy S (2013) Du financement de process par des hedge funds. Revue Lamy de Droit Civil (RLDC) 105:66–68

Faure MG, De Mot JPB (2012) Comparing third party financing of litigation and legal expenses insurance. J Law Econ Policy 8(3):743–778

Hylton KN (2012) The economics of third-party financed litigation. J Law Econ Policy 8:701. https://doi.org/10.2139/ssrn.1971229

Hylton KN (2014) Toward a regulatory framework for third-party funding of litigation. De Paul Law Rev 63(2):527–546. https://doi.org/10.2139/ssrn.2281453

Jackson J (2009) Review of civil litigation costs, final report, published with the permission of the Ministry of Justice on behalf of the controller of her Majesty's Stationery Office. Disponible online: www.judiciary.gov.uk/Resources/JCO/.../jackson-final-report-140110.pdf

Katz A (1990) The effect of frivolous lawsuits on the settlement of litigation. Int Rev Law Econ 10:3–27. https://doi.org/10.1016/0144-8188(90)90002-b

Kirstein R, Rickman N (2004) "Third Party Contingency" contracts in settlement and litigation. J Inst Theor Econ (JITE) 160(4):555–575. https://doi.org/10.2139/ssrn.427400

Lyon J (2010) Revolution in Progress: Third-Party Funding of American Litigation. Ucla Law Review 58:571–60 https://doi.org/10.2139/ssrn.2034625

McLaughlin JH (2007) Litigation funding: charting a legal and ethical course, Vanderbilt Law Review, 31: 620–621

Rubin PH (2011) Third-party financing of litigation. Northern Kentucky Law Rev 38:673–685

Schanzenbach M, Dana D (2009) How would third party financing change the face of American tort litigation?, Third party financing of litigation roundtable, Searle Center, Northwestern University Law School. Available at http://www.law.northwestern.edu/searlecenter/papers/Schanzenbach_Agency%20Costs.pdf

Veljanovski C (2012) Third-party litigation funding in Europe. J Law Econ Policy 8(3):405–449

Titling Systems

Benito Arruñada
Pompeu Fabra University and BGSE, Barcelona, Spain

Definition

Titling systems are the institutions used to enforce property rights as rights in rem and reduce the cost of transacting on them. To be effective in non-local markets, they require a registry, which produces information on claims or rights, thus allowing the judge to verify them, establish their relative priority, and solve conflicts between claimholders by adjudicating rights in rem and in personam to them. Since the judge relies on register evidence, access to registers also allows contractual parties to reduce their information asymmetry before transacting.

Introduction: The Tradeoff between Property Enforcement and Transaction Costs

Rights to land and many other assets can be enforced as property rights, *iura in rem*, claimable against the asset itself and therefore valid against all persons, *erga omnes*. These property rights are said to "run with the land," meaning that they

survive unaltered through all kinds of transactions, and transformations dealing with other rights on the same parcel of land or on a neighboring parcel. For example, the mortgagee keeps the same claim on the land even after the mortgagor sells it. Property rights oblige all people: the new owner who has purchased the land is obliged to respect both the mortgage and, in particular, the right to foreclose if the guaranteed debt is not paid. Enforcement of a property right is independent of who holds other rights on the same asset. Alternatively, rights on assets can be contract rights, enforceable against a specific person, *inter partes*. To clarify the difference between property and contract rights, consider what happens in the case of a lease of land, this being a right that in many jurisdictions may be structured as either a contract or a property right. Assume that the land is sold during the life of the lease. If the lease is a contract right, the lessee loses the right of occupation and gains instead a contract right against the lessor. However, if the lease is a property right, the lessee keeps the right of occupation. It is then the land purchaser who may have a contract right against the seller, if the sale was made free of leases. The buyer is subrogated into the seller's position. There is no change to the lease, which has run with the land from the seller to the buyer and survives intact after the sale.

When the law enforces a right as a right in rem, consent of the right holder is required for the right to be affected, that is, damaged, in any way. This requirement of consent – either real or constructive – provides precious enforcement benefits for rights on durable and immovable assets. The enforcement of contract rights, on the other hand, depends on the availability, resources, and legal status of persons, who are mobile and may become unavailable, or judgment proof when obliged to pay. For durable assets, a property right is therefore much more valuable than a contract right having the same content – that is, when the only difference between them is that the latter lacks in rem enforceability.

These enforcement benefits come at a cost, however. When multiple rights exist on an asset, transactions do not convey property rights with the promised in rem extent until all affected right holders have consented. In other words, to produce perfect property – that is, in rem – rights, some kind of explicit or implicit contracting has to take place between the transactors and each of the affected right holders, in order for the latter to give their consent. Many institutions in the field of property law are designed to make these "contracts" with affected right holders possible. Consent can be given explicitly, by private agreement, declaring to a register or in court proceedings, as well as implicitly, simply by the passing of time. Consent can also be produced at the moment the transaction takes place. Consequently, the rights resulting from the transaction will be free of uncertainty as to who the true legal right holder is and as to their precise nature. Alternatively, consent can be postponed, and the transaction then produces rights which are burdened with the survival of any property rights whose right holders have not yet consented.

In any case, without the consent of affected right holders, transactions produce a mix of property and contract rights: property – that is, in rem – effects to the extent that is compatible with the surviving property rights held by others and contract rights for the difference. The proportion of property and contract rights in the mix varies with the kind of conflicting right. In the extreme case of a fraudulent conveyance, the grantee gets only a contract right against the grantor, who is not the true legal owner. More generally, any intended property right is in fact partially contractual if an affected right holder keeps a contradictory or concurrent right against it.

Property rights thus face a trade-off with positive and negative effects (Arruñada 2012). On the one hand, they facilitate specialization by ensuring enforcement, given that right holders' consent is required to affect them. However, for the same reason, their survival after conveyance of the asset or any other transformation of rights requires costly institutions (mainly, property registers) in order to organize the process of searching, bargaining, and contracting for consent. In particular, the possibility of hidden property rights increases the information asymmetry between the conveying parties: the seller knows better

than the acquirer about hidden property rights. More generally, the need to know which conflicting property rights exist and who their holders are, and bargaining with such right holders to obtain their consent and contracting or somehow formalizing an agreement with them, all increase the costs of transforming and conveying rights. This may in turn hamper investment, trade, and specialization.

Private Titling: Privacy of Claims as the Starting Point

Under the Roman law tradition of private conveyance that was dominant in Europe until the nineteenth century, private contracts on land had in rem effects on third parties, even if they were kept secret. The baseline legal principle was that no one could deliver what they did not have (*nemo dat quod non habet*), which was closely related to the principle "first in time, first in right." So, in a double sale of land, in which an owner O sells first to buyer B_1 and later to B_2, the land belongs to B_1 because, when O sold to B_2, O was not the owner. In cases of conflict, the judge will allocate property and contract rights between both claimants (B_1 and B_2) – that is, will "establish title" – on the basis of evidence on possession and past transactions, whether or not these transactions had remained hidden.

This potential enforcement of adverse hidden rights made gathering of all relevant consents close to impossible, hindering trade, and specialization. Most transactions in land therefore gave rise, totally or partially, to contract rights, and the enforcement advantage of property rights remained unfulfilled, especially with respect to abstract rights, such as mortgages. These difficulties are clear in the functioning of the two sources of evidence traditionally used to establish title under privacy: possession and the chain of title deeds.

First, the use of possession – that is, as a first approximation, the fact of controlling the asset – as the basis for establishing property rights is a poor solution for durable assets, because for such assets it is often valuable to define multiple rights, at least separating ownership and possession. However, relying on possession to establish ownership makes it possible for possessors to fraudulently use their position to acquire ownership for themselves or to convey owners' rights to third parties. In such cases, owners will often end up holding a mere contract right, an in personam right, against the possessor committing the fraud. Understandably, under such conditions, owners will be reluctant to cede possession impersonally, for fear of losing their property. Similarly, credit will involve contractual, personal guarantees provided either by the debtor or by the lender. This is because the only way of providing some type of in rem guarantee to the lender is by transferring ownership or possession to him, thus leaving the debtor subject to the lender's moral hazard and safeguarded only by the lender's contractual guarantee, which is weaker and costly to produce.

Second, some of the problems posed by possession are solved by embodying abstract rights, such as ownership and liens, and even complementary consents in the conveying contracts, which then form a series or "chain" of title documents or deeds ("chain of deeds," for brevity) that is based on what can be labeled "documentary formalization." This evidencing of rights with the chain of deeds facilitates some degree of separation of ownership and control because it is the content and possession of deeds that provide evidence of ownership. Therefore, title experts can examine the history of transactions going back to a "root of title," which is proof of ownership in itself – either because it is an original grant from the state or, more often, because of the time that has lapsed beyond the period of prescription or the statute of limitations. However, relying on the chain of deeds also creates problems. Above all, new possibilities for error and fraudulent conveyance appear, giving rise to multiple chains of title, which leave acquirers with contract rights against the fraudulent grantor and the professionals involved in the transaction. Moreover, titles are less effective than possession in reducing the asymmetry of acquirers, as possession is observable but adverse chains of title remain hidden to the acquirer. Furthermore, acquirers remain fully unprotected against any hidden charges that are

not voluntarily contracted, such as judgment and property tax liens. Similarly, the chain of deeds also serves to enforce a security, by pledging the deeds with the lender. But this solution is also defective, as it subjects the debtor to the lender's moral hazards (the lender could impede a sale or even fraudulently sell) and causes switching costs that make mortgage subrogation difficult.

Despite these difficulties, transactions on unregistered land in England heavily relied on the chain of deeds up until the last decades of the twentieth century. Typically, ownership was proved by possession and the whole series of deeds, which was often kept by the owner's solicitor. And mortgages were formalized by pledging the deeds with the lender. Likewise, during the *ancien régime* notaries public in most civil law countries played a similar role to that of English solicitors, with the advantage that each notary office kept an archive with the original of all the titles it notarized. This gave notaries privileged access to information on the transactions that each office had authorized. However, they did not provide an effective substitute for mortgage registries, mainly because notaries' information about individual debtors was incomplete. These cases therefore illustrate a constant feature of privacy regimes: to contain fraud, private conveyancing services provided by solicitors and notaries tend to develop into professional monopolies.

Whatever the system of documentary formalization for private conveyance contracts, conveying parties will always try to contract relying on the evidence that will eventually be used by the judge to establish title. Under privacy, however, given that courts may enforce in rem rights that have remained hidden, examination of title quality is based on potentially incomplete evidence. Therefore, removal of title defects and contradictions, as well as any adjustments to the terms of the private contract, are informed only by the limited and hard-to-verify publicity provided by possession and by documentary formalization and are motivated by the risk the grantor faces when giving title warranties on a defective title. But acquirers have limited possibilities of knowing what they are buying. They, in fact, acquire residual property in rem rights plus a contract in personam right against the grantor for the difference between the in rem property rights effectively granted and what the grantor had promised to deliver.

Understandably, legal systems try to counterbalance this chronic incompleteness of property rights in their in rem dimension under a privacy regime by adopting private and public means to strengthen contract in personam rights, such as granting formal guarantees, expanding the scope of criminal sanctions, and even relying on bonding and slavery. Moreover, legal systems also provide specialized judicial procedures capable of purging title – that is, establishing which rights in rem are alive and who holds them – thus producing a public reallocation of rights that should be useful at least to solve the most complex and valuable cases.

Ceremonial Publicity and Recordation of Deeds

Whatever the palliatives applied, the costs of contracting true property – that is, in rem – rights under a regime of pure privacy are so high that modern systems of property law have abandoned privacy in an effort to lower them. At a minimum, the law induces or requires the independent publicity of contracts, which makes them verifiable, as a prerequisite for them to attain in rem effects – that is, to convey property rights and not mere contract rights. If they keep their claims private, right holders lose or risk losing in rem effects. Private contracts may create obligations among the conveying parties but do not bind third parties: all other right holders and, especially, potential future buyers and lenders. Independent publicity therefore facilitates finding out which property rights are alive and which will be affected, thus making it possible to gather consents, purge titles, and reduce information asymmetries between the conveying parties. At a maximum, in addition to requiring publicity of contracts, the law also requires a complete purge of conflicting claims for each transaction. Because this purge is supervised by a public registrar acting in a quasi-judicial capacity, the registry does

not merely provide publicity of claims but also defines and publicizes rights.

Specific laws therefore vary substantially with respect to how and when any contradiction with other property rights must be purged by obtaining the consent of the holders of these affected rights. This second contractual stage may be postponed indefinitely or may take place at the time of the private contract. In the latter case, it may be either voluntary or compulsory and total, and, if voluntary, it may also be total or partial. Moreover, jurisdictions also differ in what their registries produce, as they may either simply publicize the deeds evidencing potentially contradictory claims or certify fully purged property rights. Lastly, there is also a logical adaptation of the specific mechanisms needed to produce these outcomes in each environment, the set of rights enforced as property rights, and the adjudication rules in cases of disputed title.

Physically marking the assets is perhaps the simplest way of providing publicity of claims. The symbolic nature of marking makes it especially suitable for abstract rights, such as ownership and security interests. This explains why it has been used extensively for enforcing ownership in the absence of possession, as in valuable movables such as livestock, automobiles, and books. Another simple way of providing publicity is by using conspicuous contractual procedures. For example, ancient Roman law required a formal and public conveyance, either through a collusive proceeding before the court or through a public ceremony known as *mancipatio*. This performed a titling function, as it publicized the conveyance, but it also served to gather the consents of affected right holders and thus purged conflicting claims either at once or after a certain period. Similarly, after 1066, English conveyances followed the continental practice of delivering possession through a ritual known as livery of seisin. The publicity function was even clearer in the practices followed in some European regions, where laws mandated sophisticated procedures of publicity "before the church" and "at the gate of town walls" for rural and urban land, respectively, as well as judicial registration in some cases.

These old practices for reaching consent and purging property rights were effective because transactions mainly took place between neighbors. For neighbors, it is easy to notice announcements and public deals, especially for the kinds of rights common in rural and traditional societies, many of which were linked to family matters. It is revealing that the effect of publicity "before the church" was immediate for right holders who were present but was delayed for one year and one day for those absent. Costlier knowledge was apparently balanced with longer time, suggesting that these systems could hardly support impersonal trade.

The next logical step in the provision of publicity is to lodge private transaction documents (i.e., the title deeds) for filing in a public registry so that this evidence on property claims can then be used by the courts to verify them and allocate property, in rem, rights in case of litigation. Moreover, by making this register publicly accessible to potential acquirers, the latter can ascertain the quality of the seller's title, thus reducing their information asymmetry.

After many failed attempts, such as the Statute of Enrollments issued by Henry VIII in 1535 but never enforced, and the Massachusetts 1640 Recording Act, recordation of deeds eventually started to succeed in the nineteenth century and has been used in most of the United States, part of Canada, France, and some other countries, mostly those with a French legal background. The key for its success was to switch the priority rule, because other incentives had failed to convince people to record. Historically, recordation systems thus became effective only once courts, when deciding on a conflict with third parties, started to determine the priority of claims from the date of recording in the public office and not from the date of the deed. This means that, instead of the conventional "first in time, first in right" rule, courts adjudicate according to the rule "first to record, first in right." For instance, in terms of a double sale, the judge gives the land not to the first buyer but to the first buyer to record the purchase document.

This change in the priority rule not only protects acquirers but also avoids incomplete

recording, which hampered many of the first recordation systems. The reason is that the switch in the priority rule effectively motivates acquirers to record from fear of losing title through a second double sale or any other granting of rights (e.g., a mortgage) by the former owner to an innocent acquirer (e.g., a lender) who might record first. Consequently, all relevant evidence on property rights is available in the public records. From the point of view of third parties, the record, in principle, is complete. Other claims may not be recorded and may well be binding for the parties who have conveyed them, but these hidden claims have no effect on third parties.

The inclusiveness of the record of deeds makes it possible to assess the quality of title by having experts examine all relevant deeds, that is, only those that have been recorded, and producing "title reports." If there is sufficient demand, a whole title assurance industry will develop for examining, gathering consents, purging, and assuring title quality. This industry may take different forms. It is composed, for instance, of notaries public in France and of title insurers in many of the United States, while abstractors, attorneys, title insurance agents, and title insurance underwriters perform separate functions in other United States. Despite their different names and differing degrees of vertical integration, the industry performs similar functions in all countries, as it mainly reduces information asymmetry between the conveying parties and encourages them to voluntarily purge the title. In particular, expert search for title defects, which can thus be removed before contracting by obtaining the relevant consent. Alternatively, if not removed, the grantee will not transact or will insist on modifying the content of the private contract, reducing the price or including additional warranties, in compensation for the survival of the defect. To motivate experts' diligence and technical innovation and to spread remaining risks, a standard close to strict liability is often applied to such examination and assurance activities. Consequently, experts are strongly motivated to find any defects on the title and a substantial part of the remaining title risk is reallocated from acquirers to title experts and their insurers.

Moreover, as under the privacy regime, both contractual and judicial procedures are used to remove title defects. Compared to privacy, deed recordation provides more possibilities for contracting the removal of defects, because defects are better known to buyers and insurers. The identification of right holders also gives greater security to the summary judicial hearings that serve to identify possible adverse claims and publicly reallocate in rem rights. These summary hearings continue to exist today in, for example, the French judicial purge and the US "quiet title" suit. In addition to purging titles directly, the existence of such a court-ordered purging possibility also reduces bargaining costs indirectly by encouraging recalcitrant claimants to reach private agreements.

However, the recording office accepts all deeds respecting certain formal requirements (mainly, the date of the contract and the names of the conveying parties), whatever their legality and their collision with preexisting property rights. In fact, the recording office is often obliged by law to file all documents fulfilling a set of formal requirements, regardless of their legal status. The public record may therefore contain three kinds of deed. First, those resulting from private transactions made without previous examination. Second, those granted after an examination but without having all defects removed. Finally, those that define purged and noncontradictory property rights.

Transactors who record clouded titles therefore produce a negative externality for all future transactors. When examining the title of a parcel, experts do not know a priori which kinds of deed are recorded concerning it. Therefore, for each transaction, they will have to examine all relevant deeds dealing with that parcel, even those which may have been perfectly purged in previous transactions. The cost of this repeated examining of deeds can be reduced with proper organization of the registry. In the short run, the easiest way to organize the information is by relying on indexes of grantors and grantees to locate the chain of transactions for a given parcel. However, this method is subject to errors, such as those caused by identical names and misspellings. This

explains the steps taken, for instance, in 1955 to create the *fichier immobilier* in the French Registry and to forbid recording a deed if the grantor's title is not recorded. Another way of reducing costs when public records are poorly organized is to build privately owned, indexed databases (known in the United States as "title plants"). These plants replicate public records in a more complete, reliable, and accessible manner by transferring and abstracting relevant documents lodged at the public registries and building tract indexes to easily locate the relevant information for each land parcel.

Registration of Rights

Registration of rights (hereafter referred to as "registration," and often confusingly called "title registration") goes one crucial step further than recordation of deeds: instead of providing information about claims, it defines the rights (what, in jurisdictions with Torrens registries, is often referred to as "title by registration"). To do this, it requires a mandatory purge of claims before registering the rights. As in deed recordation, claims stemming from private transactions gain priority when transaction documents are first lodged with the registry. They are then subject, however, to substantive review by the registrar in order to detect any potential conflict that might damage other property rights. New and reallocated rights are registered only when the registrar determines that the intended transaction does not affect any other property right or that the holders of these affected rights have consented. When these conditions are met, the change in rights caused by the transaction is registered, antedating the effects of registration to the lodging date. (In a sense, any registry of rights thus contains a recording of deeds: its "lodgment" or "presentment" book is a temporary record of claims.) Otherwise, when the consent of an affected right holder is lacking, registration is denied, and the conveying parties have to obtain the consents relevant to the originally intended transaction, restructure it to avoid damaging other rights, or desist.

Registration aims to eliminate all uncertainties and information asymmetries, as information in the register is simplified in parallel with the purge of rights. Ideally, rights defined in each new contract are registered together with all surviving rights on the same parcel of land. Extinguished rights are removed or deleted, making it easy to know which are the valid rights. Production of information is a key element. As pointed out by Baird and Jackson, "in a world where information is not perfect, we can protect a later owner's interest fully, or we can protect the earlier owner's interest fully. But we cannot do both" (1984, p. 300). The assertion is accurate but the assumption is crucial: registration intends to produce perfect information and thus protect both the earlier and the later owners. Its goal is to abide by three principles traditionally deemed desirable for a titling system, according to which: (1) the register reflects the reality of property rights, so that potential transactors do not need to look out of the register ("mirror principle"); (2) the register reflects only valid rights, so that transactors do not need to perform a title search in the chain of title ("curtain principle"); and (3) losses caused by a registry's failure are indemnified ("assurance, insurance, or guarantee principle").

To the extent that these three principles are achieved and given that any contradictions are purged before registration, the registry is able to provide "conclusive," "indefeasible" title, meaning that a good faith third party "for value" (i.e., one who pays for the property rather than receiving it as a gift) acquires a property right if the acquisition is based on the information provided by the registry. If the seller's right is later shown to be defective, the buyer keeps the property – that is, in rem – right and the original owner gets contract in personam rights against the seller and the registry. When functioning correctly, the register is thus able to provide potential transactors with a complete and updated account of the in rem rights alive on each parcel of land. Given the enforcement advantage of in rem rights, this accounting amounts to commoditizing the legal attributes of rights on real property, which makes impersonal trade much easier.

Furthermore, registration interferes little with private property, as registry intervention focuses

on the timing and completeness of the reallocation of rights implicit in any purge. Registration is controlled by registrars, but ultimate decisions are made by right holders by giving their consent. Privacy and recordation allow conveying parties more discretion on timing and heavier reliance on privately produced information. They therefore seem to rely more on private decisions, but this perception is deceptive because even recorded titles are in fact mere claims. They retain a higher contractual content, given the survival of conflicting claims in rem. Additional intervention by the court, also subject to the possibility of allocation failure, would be required to transform them into property rights at in rem level equivalent to that provided by registration. In sum, as compared to recordation, it is useful to see registration as a quasi-judicial step, which in other titling systems is also necessary to reach full in rem enforcement. This similar degree of public involvement helps explain why both registration and recordation have taken root in countries with different legal traditions and why there is little correlation between titling systems and legal traditions.

Cross-References

- ▶ Coase and Property Rights
- ▶ Development and Property Rights
- ▶ Emissions Trading
- ▶ Externalities
- ▶ Good Faith
- ▶ Informal Sector
- ▶ Institutional Change
- ▶ Lex Mercatoria
- ▶ Limits of Contracts
- ▶ Market Failure: Analysis
- ▶ Public Enforcement
- ▶ Transaction Costs

References

Arruñada B (2012) Institutional foundations of impersonal exchange: the theory and policy of contractual registries. University of Chicago Press, Chicago

Baird DG, Jackson TH (1984) Information, uncertainty, and the transfer of property. J Leg Stud 13:299–320

Further Reading

Arruñada B (2003) Property enforcement as organized consent. J Law Econ Org 19:401–444

Arruñada B (2011) Property titling and conveyancing. In: Ayotte K, Smith HE (eds) Research handbook on the economics of property law. Edward Elgar, Cheltenham, pp 237–256

Arruñada B (2015) The titling role of possession. In Chang Y (ed) The law and economics of possession. Cambridge University Press, Cambridge

Arruñada B, Garoupa N (2005) The choice of titling system in land. J Law Econ 48:709–727

Deininger K, Feder G (2009) Land registration, governance, and development: evidence and implications for policy. World Bank Res Obs 24:233–266

Ellickson RC, Thorland CD (1995) Ancient land law: Mesopotamia, Egypt, Israel. Chicago-Kent Law Rev 71:321–411

Hansmann H, Kraakman R (2002) Property, contract, and verification: the numerus clausus problem and the divisibility of rights. J Leg Stud 31:S373–S420

Merrill TW, Smith HE (2000) Optimal standardization in the law of property: the numerus clausus principle. Yale Law J 110:1–70

Merrill TW, Smith HE (2001) The property/contract interface. Columbia Law Rev 101:773–852

Merrill TW, Smith HE (2001) What happened to property in law and economics? Yale Law J 111:357–398

Tort Damages

Louis Visscher
Rotterdam Institute of Law and Economics (RILE), Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Definition

The amount of monetary compensation a tortfeasor has to pay to the plaintiff(s) in a tort case when he is found liable.

Introduction

Liability rules in tort law determine when a tortfeasor is liable. The rules of tort damages

determine the amount the liable tortfeasor subsequently has to pay. Together these two bodies of law therefore determine in which situations the tortfeasor has to pay how many damages to the plaintiff(s). Therefore, the behavioral incentives provided by tort law depend on both sets of rules. In this entry, the most important insights from the economic analysis of tort damages are presented. The limited space does not allow a full discussion of all possible complications (see Visscher 2009 for a more complete overview), but where relevant such complications are briefly indicated. It is also not possible to include a full discussion of the liability rules themselves, but in the remainder of this introduction, a very brief account of this topic is provided.

The economic analysis assumes that actors are incentivized to take precautionary measures by the threat of liability (tortfeasors) or the prospect of having to bear one's own losses (victims). Such measures can consist of taking care and/or adapting the activity level. Optimal measures are taken when the marginal costs equal the marginal benefits (see, e.g., Schäfer and Ott 2005; Cooter and Ulen 2012). The social benefits of precautionary measures consist of the reduction in expected accident losses (i.e., the probability of an accident multiplied by the magnitude of the losses if an accident happens). The private benefits of a party consist of the reduction in the losses they themselves have to bear. In order to provide socially desirable incentives, tort damages therefore should equal the social losses, which in principle calls for full compensation (restitutio in integrum), and negligence should be defined as taking less than socially optimal care.

An essential difference between strict liability and negligence is that under the first rule the tortfeasor is liable irrespective of his care level and under the latter rule he is only liable if he took less than due care. This implies that whereas under strict liability tort damages indeed should be full, under negligence tort damages can be lower, because they only have to make taking due care (and hence escaping liability) more attractive than taking low care (and hence being liable). However, if a negligent tortfeasor is only held liable for the losses which are caused *by his negligence*, this difference disappears and also negligence requires full compensation.

In bilateral accident situations, where both the tortfeasor and the victim can influence the accident probability, both parties should receive incentives for optimal precautions. Negligence provides optimal care incentives to both parties (the tortfeasor takes optimal care to avoid liability and the victim to minimize the costs he himself has to bear), but strict liability needs a defense of comparative or contributory negligence to achieve this. No liability rule can give optimal activity incentives to both parties, because only the party who ultimately bears the accident losses will choose the correct activity level.

Pecuniary losses are either monetary losses or losses of replaceable goods, where the replacement costs are a good measure of the losses. Nonpecuniary losses consist of damage to irreplaceable things such as family portraits but also health and emotional well-being (Shavell 2004, p. 242). For pecuniary losses, the concept of full compensation is relevant, because damages can make the victim indifferent between the situation without the tort on the one hand and the situation with the tort and with damages on the other hand (also see section "What is Full Compensation?"). With nonpecuniary losses, it is problematic that money cannot truly compensate such losses, that they cannot be observed directly, and that there is no market so that one needs to apply indirect methods of assessing them (see section "Nonpecuniary Losses"). In order to provide the correct incentives to the injurer, tort damages should equal the sum of pecuniary and nonpecuniary losses (Shavell 2004, p. 242).

What is Full Compensation?

In many European countries, the *difference hypothesis* of Mommsen is important in assessing the losses which the liable tortfeasor has to compensate. In this approach, the loss consists of the difference between the wealth as it is at a certain moment and the wealth as it would have been at the relevant moment if the loss causing event would not have happened. In economic terms,

this can be expressed as follows: before the accident, the victim had a certain level of utility. After the accident, he has a lower level of utility, and the difference in both utility levels is his loss. Even though Mommsen defined the loss as a difference between two wealth levels, in many jurisdictions also nonpecuniary losses are compensable. Again in economic terms, the decrease in utility can not only be caused by a reduction in wealth but also by a reduction in other sources of utility, which may have a nonpecuniary nature.

As explained in the Introduction, in order to provide correct incentives, tort damages in principle should fully compensate the victim for his losses (Posner 2003, p. 192; Cooter and Ulen 2012; Shavell 2004, p. 236) and should be based on the social losses. This way, the tortfeasor internalizes the negative externality he has created. By receiving an adequate amount of money, the victim can be brought back to his original utility level, or better, to the utility level he would have reached without the tort. Under negligence, if the tortfeasor would have taken due care, the victim also would run a certain risk of being harmed, so that he should not be brought to a position in which he would run no risk at all. This implies that the victim should not be compensated for losses which also would have happened at due care (Van Wijck and Winters 2001).

Several remarks can be made regarding the principle of full compensation:

- If victims can mitigate their losses by taking measures which cost less than the reduction in accident losses they yield, they should be incentivized to do so. Limiting damages to optimally mitigated losses plus mitigation costs provides the proper incentives (Shavell 2004, p. 248ff).
- If repair is more expensive than replacement, damages should be based on replacement. If repair or replacement is not (completely) possible, monetary damages should be high enough to bring the victim back to his original utility curve.
- Costs which the victim would have incurred anyway should not be compensated (e.g., gasoline costs of a rental car which is needed during repair of the victim's car).
- Damages should take interest, taxes, and inflation into account. For future losses, there is a preference for lump sum rather than periodic contingent payments. A victim who prefers periodic payments can convert the lump-sum payment into periodic payments, lump-sum payment avoids the uncertainties of periodic contingent payment as well as possible moral hazard on the side of the victim, and the victim is not confronted with ongoing monitoring to assess the extent of the losses which would be required under periodic contingent payment (Rea 1981, p. 132).
- The injurer has to compensate the losses of the victim, also if they are higher than normal. Such losses should not be limited on the basis of foreseeability or adequacy but rather the tortfeasor should take the victim as he finds him. After all, limiting damages if they are higher than normal would result, on average, in too few incentives (Shavell 2004, p. 239).
- Judicial moderation and limitation of damages are generally critically assessed by law and economics scholars, because they result in less than full compensation. The often-heard argument of uninsurability of the full losses is not convincing, because it is not primarily the magnitude of losses that results in uninsurability but uncertainty regarding the accident probability. Restricting liability does not solve that issue and might even result in higher losses by diminishing tort law's care incentives (Cooter and Ulen 2012; Shavell 2004, p. 230ff). Restricted liability may induce a potential tortfeasor to engage in a socially desirable activity which he, due to risk aversion, might not engage in otherwise. But it could also induce a potential tortfeasor to engage in an activity which is socially *un*desirable due to the high costs, of which he now only has to bear a fraction.
- One could base damages not on the true losses in individual cases but on the average loss. Such an abstract, objective method of damage assessment instead of a concrete, subjective method can greatly reduce the administrative

costs of tort law. For example, tort damages for a damaged car can be based on the costs of repair by a competent mechanic, even if in the actual case the victim did not have the car repaired or did it himself. The reduction in administrative costs in such often-occurring losses outweighs the possible disadvantage of less fine-tuned deterrence incentives. It is even the question if there is a reduction in deterrent incentives in the first place, because a better damage assessment ex post does not result in better incentives if the tortfeasor ex ante does not know the exact losses he may cause. As long as the abstract method assesses the average damages correctly, the injurer receives the correct incentives (Kaplow and Shavell 1996). If losses are systematically over (under) estimated, the tortfeasor is generally assumed to choose a too high (low) care level and a too low (high) activity level.

- Harm to the victim generally speaking is a better basis for tort damages than gain to the injurer. In the latter case, if courts would underestimate the gain, the incentives would be inadequate (Polinsky and Shavell 1994, p. 431ff). Furthermore, basing damages on the loss results in internalization of the externality and allows an activity which yields more benefits than costs to be continued. However, if harm to the victim is difficult to assess, e.g., due to its subjective nature, basing damages on the gain may provide better deterrent incentives. Also if one wants to fully deter the behavior, rather than "merely" inducing the tortfeasor to take optimal precautionary measures, basing damages on gain is preferable.
- If litigation costs are taken into account, optimal damages may differ from full compensation of harm (Polinsky and Shavell 2014).

Nonpecuniary Losses

As explained in the Introduction, nonpecuniary losses are difficult to assess because they cannot be observed directly (see e.g. Arlen 2013, p. 439ff). This could induce victims to overstate such losses, and therefore it is sometimes suggested not to compensate them at all when they are likely to be small. This would greatly reduce administrative costs and only have a limited impact on deterrence. Adams provides a different argument why such losses should not be compensated. Victims who have to bear these losses themselves receive desirable behavioral incentives (Adams 1989, p. 215ff). The problem with this view is, of course, that tortfeasors then receive too few incentives because they only compensate part of the losses. If nonpecuniary losses are too large to ignore, one could use tables or formulas to determine damages, which is a form of an abstract method as discussed in section "What is Full Compensation?" In order to avoid structural over- or undercompensation, it is important to try to find a method which assesses such losses correctly on average (if at all possible). This idea will be further developed below.

Given that nonpecuniary losses cannot be observed directly, an indirect way to assess them is required. The most well-known indirect method in this respect is provided by the *insurance theory*. This approach argues that a victim should only receive compensation for losses against which he is willing to insure himself. Otherwise, tort law would force coverage upon him which he does not want, but for which he might have to pay via higher prices for goods or services produced by the liable tortfeasor. According to the insurance theory, nonpecuniary losses do not increase marginal utility of wealth, and hence rational victims are not willing to insure against such losses, because the premium would cost more utility than the expected coverage yields (Friedman 2000, p. 95ff; Shavell 2004, p. 270ff). The fact that in practice there is no demand for such insurance is regarded as corroboration of the line of reasoning. The insurance theory therefore claims that victims should not receive compensation for nonpecuniary losses. The injurer, however, should pay for them, so that liability should be decoupled from compensation (Geistfeld 1995, p. 799ff).

Various critiques regarding the insurance theory have been expressed. First, even if marginal utility of wealth would not increase, potential victims might still want to insure against such

losses because the money they would receive after suffering these losses would mitigate the decrease in the *level* of utility, irrespective of the *marginal* utility of wealth at that point. Second, the fact that there is no demand for insurance against nonpecuniary losses may be caused by imperfect information regarding the extent of nonpecuniary losses, the probability of their occurrence, and the compensation required in respect of them or by countervailing social norms in the form of societal hostility to putting a price on pain and sorrow and legal restrictions such as the indemnity principle (Croley and Hanson 1995, p. 1845ff). Third, the argument that nonpecuniary losses do not result in an increased marginal utility of wealth may be flawed to start with. Empirical research which suggests that marginal utility of wealth decreases after suffering nonpecuniary losses is mostly based on personal injury situations where nondisabled people are asked how they think personal injuries would affect them if they would suffer such losses. It is doubtful whether nondisabled people can accurately assess the impact of such injuries on the marginal utility of money, and insights regarding adaptation suggest that this research results in an underestimation of marginal utility which may in fact increase after the nonpecuniary loss is suffered (Pryor 1993, p. 116ff).

But even if marginal utility would not increase, and even if people would not self-insure against nonpecuniary losses, it is questionable whether this should lead to the conclusion that victims should not receive compensation for such losses. Insurance decisions provide information on the degree of risk aversion of the actor involved but not on his willingness to pay to avoid the loss. If the victim would be willing to forego resources ex ante in order to avoid the nonpecuniary loss or at least to lower the chances of suffering them, this would imply that he does regard these losses as undesirable and that they do lower his utility level. The resources the victim himself is willing to spend on accident avoidance therefore provide information of how the victim himself assesses the nonpecuniary loss. This is a better indirect way of assessing damages for nonpecuniary losses than the insurance decision, which – again – regards risk aversion rather than loss assessment. By basing damages on the resources victims themselves would have been willing to spend on accident avoidance, the victims are not overcompensated against their will, nor are potential injurers overdeterred because, in essence, they bear the costs of avoidance measures which were worthwhile ex ante.

For fatal accidents, the so-called *Value of a Statistical Life* (VSL) provides information on the willingness to pay to lower the chance of fatal accidents. The VSL is derived from all kinds of decisions people take which affect health and safety, such as wearing a helmet when riding a (motor)bike, installing a smoke detector at home, asking a premium for dangerous work conditions, etc. Such decisions contain an implicit trade-off between money and risks. If, for example, 1,000 people are each willing to spend € 1,000 on a measure which reduces the probability of a fatal accident by one pro mille, one statistical life is saved by spending € 1,000,000. The VSL in this example then is at least € 1,000,000 because this population was willing to spend at least that amount. In 2004, Sunstein assessed the VSL at about $6.1 million (Sunstein 2004, p. 205; Posner and Sunstein 2005, p. 563). Expressed in 2014 euros this would amount to about €6.4 million. This American VSL is comparable to that in other developed countries (Viscusi and Aldy 2003, pp. 24, 35, and 63). If one wants to distinguish between victims on the basis of their age, one should use the *Value of a Statistical Life-Year* (VSLY) instead. Posner and Sunstein argue that damages for fatal accident which are based on the VSLY should be around $6 million or higher and including the emotional loss of surviving relatives could result in an increase of several millions of dollars as well (Posner and Sunstein 2005, p. 586ff). These amounts are much higher than currently awarded in European countries, where damages are often limited to funeral costs and loss of maintenance. This implies that from an economic point of view, damages for fatal accidents are structurally undercompensated.

Schäfer and Ott argue that for nonfatal injuries, damages should be "some fraction" of the value attached to the willingness to pay to prevent death (Schäfer and Ott 2005, p. 373), but they do not

suggest how the appropriate fractions could be determined. The concept of the *Quality-Adjusted Life-Year* (QALY) could play a good role here (Miller 2000; Karapanou and Visscher 2010). QALYs express the value of living 1 year in a certain health condition, which is a proxy of the quality of life during that period. By combining information regarding the QALY value of that health condition and the duration thereof on the one hand and a monetary value of a QALY on the other hand, it is possible to express in monetary terms the loss of utility due to suffering personal injuries. QALYs are used to evaluate health programs and medical treatments and techniques and hence express how much society is willing to prevent or cure the health conditions under consideration. They therefore can provide information on the societal willingness to pay to avoid or cure such losses and are therefore suitable as a basis for the assessment of pain and suffering damages for personal injuries. It turns out that applying this method, even with conservative monetary values, results in higher pain and suffering damages than currently awarded in most European countries, so that here as well, currently these losses are structurally undercompensated.

Pure Economic Loss

In the Introduction, it became clear that tort damages should in principle be based on the social loss the tortfeasor has caused. In cases of pure economic loss, the private losses of the victim often do not coincide with the social loss. The private losses might be offset by private gains elsewhere, so that the tort resulted in a redistribution of social welfare rather than in a reduction thereof. If firm A cannot produce due to a power supply interruption caused by a tort, but firm B now produces and sells more, there might not be a social loss at all, and tort liability then makes no economic sense. This redistribution argument is regarded as a reason not to include pure economic loss in tort damages (Bishop 1982; Gómez and Ruiz 2004; Dari-Mattiacci and Schäfer 2007, p. 10). However, several reasons exist why there might be a social loss after all.

First, the products of firm B might not be a perfect substitute so that there is a loss of consumer surplus.

Second, firm B had to have overcapacity in order to meet the additional demand, and overcapacity in itself can be regarded as inefficient. Compensating pure economic loss provides better behavioral incentives to potential tortfeasors, resulting in fewer accidents and hence a reduced need for overcapacity (Rizzo 1982, p. 202).

Third, there will be many cases of pure economic loss where there is not only a redistribution of wealth but also a decrease in social welfare. For example, if an accountant negligently approves the balance sheet of a firm, besides the redistribution of welfare between buyers and sellers of overpriced stock, such faults may decrease trust in information provided by accountants and induce parties to do additional research and result in suboptimal investment decisions. These are all examples of social losses.

Gómez and Ruiz present two other reasons why pure economic loss should sometimes be included in tort damages. In situations where the victim in essence "insured" someone else against losses by covering them, it is desirable that he can recover his pure economic loss from the tortfeasor due to the same reasons why subrogation in real insurance contracts is desirable. An example of this is an employer who pays wages to his employee also during the period where the latter cannot work because he is injured in a tort. It is also possible that a third party suffers pure economic loss due to breach of contract. The third party might not be able to recover these losses under contract law because he is no party to the contract. Tort law could then serve as a surrogate for contractual liability (Gómez and Ruiz 2004).

Punitive Damages

Punitive damages are damages which exceed compensatory damages and are used to punish wrongdoers in cases of, e.g., malice, gross

negligence, or intent. Continental European jurisdictions generally do not award punitive damages. In debates regarding punitive damages, a fear for "American situations" in which enormous amounts of damages are awarded is often expressed. However, such large amounts are often reduced by courts, and empirical research shows that punitive damages are not often awarded and generally speaking are not very high, and they are correlated with compensatory damages (Cohen and Harbacek 2011; Vidmar and Wolfe 2009, p. 189, However, also see Polinsky 1997). The "American situations" are therefore more myth than reality, and fear for them should not frustrate the debate on introducing punitive damages in continental Europe. This holds even stronger now there are good economic arguments in favor of punitive damages (also see Arlen 2013, p. 488ff).

First, punitive damages can ameliorate the problem that the probability that a tortfeasor will actually be held liable is below 100%. By increasing damages, the *expected* damages can again equal social harm. For example, if the probability of being held liable for causing a loss of 100 is 25%, but total damages (consisting of compensatory and punitive damages) are 400, the expected liability is again 100 (25%*400) (Cooter 1989; Polinsky and Shavell 1998). This idea is consistent with the fact that punitive damages are often awarded in cases of intentional torts, because there the injurer might try to avoid being caught.

Second, if compensatory damages do not cover all (types of) losses of the victim, damages are too low from an economic viewpoint because they do not cover the full externality. Such undercompensatory damages can result from losses which are difficult to assess, e.g., because they are nonpecuniary or subjective. In principle, punitive damages could tackle this problem. The fact that in cases of defamation (with indeed subjective and nonpecuniary losses) punitive damages are often possible fits this idea. However, in order to assess the correct magnitude of punitive damages, one has to assess the extent of undercompensation of normal damages, and if one would be able to assess that, one could better increase compensatory damages by this amount (Polinsky and Shavell 1998, p. 940).

Third, if the utility a tortfeasor yields from his activity is regarded as socially illicit so that we do not want to include it into the social weighing of costs and benefits, compensatory damages may not adequately deter the undesirable behavior because the tortfeasor still yields his utility. Punitive damages could result in more deterrent incentives (Cooter 1989, p. 79ff; Shavell 2004, p. 245). The value of this third rationale is limited, because (1) many of the acts which yield socially illicit utility will be subject to criminal liability and (2) labeling certain benefits as socially illicit *assumes* the conclusion that these acts are socially undesirable rather than that this is *proven* on the basis of a weighing of costs and benefits (Friedman 2000, p. 229ff).

Fourth, the threat of punitive damages could induce a potential tortfeasor to seek a voluntary transaction with the potential victim, rather than to commit the tort. Law and economics scholars generally prefer such voluntary transactions because then also subjective elements in the valuation are included in the decision whether or not to transfer an entitlement, whereas tort damages do not include these elements, so that they might not fully compensate the victim. In cases where a voluntary transaction was possible, the potential tortfeasor should be incentivized not to take away the entitlement without consent and to pay the objective damages afterward, exactly because of the possible subjective elements. Punitive damages can help by making the involuntary transfer more expensive (Shavell 2004, p. 245ff).

Fifth, victims who start a tort claim in essence serve the social goal of deterrence, yet they bear all the costs. If the costs outweigh the expected private benefits, they may stay rationally apathetic. Punitive damages, such as treble damages in antitrust cases, may ameliorate this situation.

Sixth and final, in as far as independent value is attached to the punishment goal (the desire of individuals to see blameworthy parties appropriately punished), punitive damages can ensure that more reprehensible behavior results in a more severe sanction (Polinsky and Shavell 1998,

p. 948ff). This is consistent with the fact that punitive damages are often connected to malice, gross negligence, or intent.

Conclusion

As a starting point, tort damages should fully encompass the externality caused by the tortfeasor. Under negligence, lower damages are possible as long as they make taking optimal care more attractive than being negligent. Victims should receive incentives to take optimal precautionary and mitigation measures. Tort damages should encompass both pecuniary and nonpecuniary losses, unless the administrative costs of assessing the latter outweigh the behavioral benefits of including them. In principle, especially for difficult classes of losses, an objective abstract assessment of damages suffices, as long as systematic over- or undercompensation is avoided. It seems that actual nonpecuniary damages for fatal accidents and personal injuries are (much) lower than economically desirable. Pure economic loss should only be compensated if it also entails a social loss. Punitive damages serve several economic goals, and the European debate should not be shut down by a simple reference to "American situations," which are more myth than reality.

References

Adams M (1989) Warum kein Ersatz von Nichtvermögensschäden? In: Ott C, Schäfer HB (eds) Allokationseffizienz in der Rechtsordnung. Springer, Berlin, pp 210–217

Arlen J (ed) (2013) Research handbook on the economics of Torts. Edward Elgar, Cheltenham

Bishop W (1982) Economic loss in Tort. Oxford J Legal Stud 2:1–29

Cohen TH, Harbacek K (2011) Punitive damage awards in state courts, 2005. US Department of Justice Special Report March 2011

Cooter RD (1989) Punitive damages for deterrence: when and how much? Ala Law Rev 40:1143–1196

Cooter RD, Ulen TS (2012) Law and economics, 6th edn. Addison Wesley, Boston

Croley SP, Hanson JD (1995) The nonpecuniary costs of accidents: pain and suffering damages in Tort law. Harv Law Rev 108:1785–1917

Dari-Mattiacci G, Schäfer HB (2007) The core of pure economic loss. Int Rev Law Econ 27:8–28

Friedman DD (2000) Law's order. What economics has to do with law and why it matters. Princeton University Press, Princeton

Geistfeld M (1995) Placing a price on pain and suffering: a method for helping juries determine Tort damages for nonmonetary injuries. Calif Law Rev 83:773–852

Gómez F, Ruiz JA (2004) The plural – and misleading – notion of economic loss in Tort: a law and economics perspective. Z Eur Priv 12:908–931

Kaplow L, Shavell S (1996) Accuracy in the assessment of damages. J Law Econ 39:191–210

Karapanou V, Visscher LT (2010) Towards a better assessment of pain and suffering damages. J Eur Tort Law 1:48–74

Miller TR (2000) Valuing nonfatal quality of life losses with quality-adjusted life years: the health economist's meow. J Forensic Econ 13:145–167

Polinsky AM (1997) Are punitive damages really insignificant, predictable and rational? A comment on Eisenberg et al. J Legal Stud 26:663–677

Polinsky AM, Shavell S (1994) Should liability be based on the harm to the victim or the gain to the injurer? J Law Econ Organ 10:427–437

Polinsky AM, Shavell S (1998) Punitive damages: an economic analysis. Harv Law Rev 111:869–962

Polinsky AM, Shavell S (2014) Costly litigation and optimal damages. Int Rev Law Econ 37:86–89

Posner RA (2003) Economic analysis of law, 6th edn. Aspen Publishers, New York

Posner EA, Sunstein CR (2005) Dollars and death. Univ Chic Law Rev 72:537–589

Pryor ES (1993) The Tort law debate, efficiency, and the kingdom of ill: a critique of the insurance theory of compensation. Virg Law Rev 79:91–152

Rea S (1981) Lump sum versus periodic damage awards. J Legal Stud 10:131–154

Rizzo MJ (1982) The economic loss problem: a comment on bishop. Oxf J Legal Stud 2:197–206

Schäfer HB, Ott C (2005) Lehrbuch der ökonomischen Analyse des Zivilrechts, 4th edn. Springer, Berlin

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

Sunstein CR (2004) Lives, life-years, and willingness to pay. Columbia Law Rev 104:205–252

Van Wijck P, Winters JK (2001) The principle of full compensation in Tort law. Eur J Law Econ 11: 319–332

Vidmar N, Wolfe MW (2009) Punitive damages. Ann Rev Law Soc Sci 5:179–199

Viscusi WK, Aldy JE (2003) The value of a statistical life: a critical review of market estimates throughout the world. J Risk Uncertain 27:5–76

Visscher LT (2009) Tort damages. In: Faure MG (ed) Tort law and economics, vol I, 2nd edn, Encyclopedia of law and economics. Edward Elgar, Cheltenham, pp 153–200

Tourism

Marianna Succurro
Department of Economics, Statistics and Finance, University of Calabria, Rende (CS), Italy

Abstract

The term tourism indicates all the heterogeneous activities and services referring to the temporary transfer of people from the habitual residence to other destinations for diverse reasons. It is a social, cultural, and economic phenomenon which is very important, and, in some cases vital, for many countries. Over the past decades, tourism has experienced continued expansion and diversification, becoming one of the largest and fastest-growing economic sectors in the world. It generates also direct effects on the social, cultural and educational sectors of national societies and on their international relations. Due to these multiple impacts, there is a need for a holistic approach to tourism analysis, development, management and monitoring in order to formulate and implement national and local tourism policies and international agreements.

Definition

The term tourism indicates all the heterogeneous activities and services referring to the temporary transfer of people from the habitual residence to other destinations for leisure, amusement, entertainment, culture, medical treatment, business, sport, and other purposes.

These people are called visitors (which may be either tourists or excursionists, residents or nonresidents), and tourism has to do with their activities, some of which imply tourism expenditure.

In 1936, the League of Nations defined a foreign tourist as "someone traveling abroad for at least 24 h." Its successor the United Nations amended this definition in 1945, by including a maximum stay of 6 months.

Tourism and its Multiple Impacts

Tourism is a social, cultural, and economic phenomenon which is very important, and, in some cases vital, for many countries.

Many researchers reckon it is a key driver of socioeconomic progress through export revenues, the creation of jobs and enterprises, and infrastructure development. As such, tourism has implications on the economy, on the natural and built environment, on the local population at the destination, and on the tourists themselves.

In the *Manila Declaration on World Tourism of 1980*, tourism is recognized as "an activity essential to the life of nations because of its direct effects on the social, cultural, educational, and economic sectors of national societies and on their international relations." Due to these multiple impacts, there is a need for a holistic approach to tourism analysis, development, management, and monitoring in order to formulate and implement national and local tourism policies and international agreements.

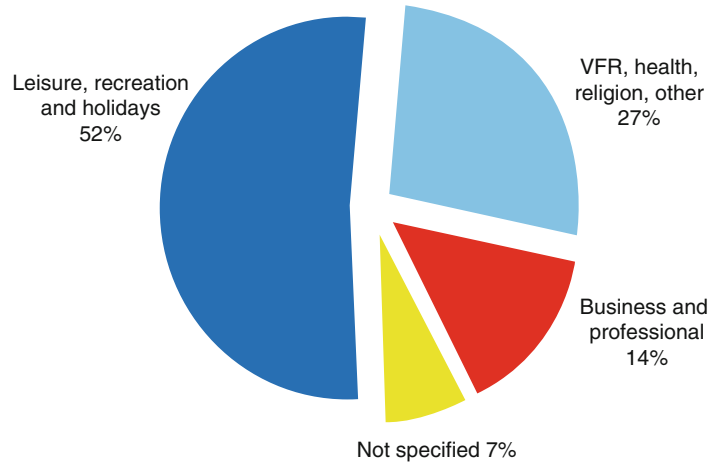
The Increasing Importance of Tourism in the World as Consumer Good

The origins of modern tourism can be traced back to the traditional trip of Europe undertaken by mainly upper-class European young men. The custom flourishes from about 1660 until the advent of large-scale rail transit in the 1840s, usually associated with a standard itinerary.

Mass tourism develops with the improvements in infrastructures and technology, allowing the transport of large numbers of people in a short space of time to places of leisure interest, so that greater numbers of people can begin to enjoy the benefits of leisure time.

Over the past six decades, indeed, tourism has experienced continued expansion and diversification, becoming one of the largest and fastest-growing economic sectors in the world. Many new destinations have emerged in addition to the traditional favorites of Europe and North America.

Despite occasional shocks, international tourist arrivals have shown virtually uninterrupted



Tourism, Fig. 1 Inbound tourism by purpose of visit, 2013 (share) (Source: World Tourism Organization)

growth from 25 million in 1950 to 278 million in 1980, 528 million in 1995, and a record of 1087 million arrivals in 2013 (UNWTO 2015).

Nowadays, tourism is considered a popular global leisure activity.

Most travel is for leisure purposes (52%), followed by health and religion purposes (27%), business (14%), and others (Fig. 1).

According to the *UNWTO*'s long-term forecast *Tourism Towards 2030*, international tourist arrivals worldwide are expected to increase by 3.3% a year from 2014 to 2030 to reach 1.8 billion by 2030.

The developments in technology and in transport infrastructure, such as low-cost airlines and more accessible airports, have also made many types of tourism more affordable. Internet sales of tourist services have facilitated changes in lifestyle. Some sites have started to offer dynamic packaging, in which an inclusive price is offered for an appropriate package requested by the customer (see paragraph 4.1 on ICT and tourism distribution).

Tourism Data and Statistics

Secondary Data and Statistics

Worldwide Tourism Statistics

The United Nations World Tourism Organization (UNWTO) *Tourism Highlights* is a very useful source for a long-term outlook on tourism. The UNWTO is the United Nations' agency responsible for the promotion of responsible, sustainable, and universally accessible tourism. As the leading international organization in the field of tourism, UNWTO promotes tourism as a driver of economic growth, inclusive development, and environmental sustainability and offers leadership and support to the sector in advancing knowledge and tourism policies worldwide. It is the leading organization collecting and disseminating the most up-to-date and comprehensive tourism data, short- and long-term forecasts, and knowledge on specific and source markets; thus it serves as a global forum for tourism policy and a source of tourism know-how.

The *UNWTO Tourism Towards 2030* is UNWTO's long-term outlook and assessment of the development of tourism for the two decades from 2010 to 2030, and it has become a worldwide reference for international tourism forecasts.

Understanding, for each country, where its inbound tourism is generated is essential for analyzing international tourism flows and devising marketing strategies, such as those related to the positioning of national markets abroad.

UNWTO's main dataset and publication on annual tourism statistics include:

– *Yearbook of Tourism Statistics*, which, deriving from the most comprehensive statistical database available on the tourism sector, focuses on data related to inbound tourism

(total arrivals and overnight stays), broken down by the country of origin
– *Compendium of Tourism Statistics*

Complete data and statistics for European countries are also available from the Eurostat database.

National Institutes of Statistics
Additional statistics can be found in publications on tourism made available by the National Institute of Statistics in each country.

The United Nations General Assembly has endorsed the *Fundamental Principles of Official Statistics* (last 29 January 2014). These principles are considered a basic framework which all statistical activities developed by national and international organizations must follow in recognizing official statistics as a public good.

Tourism Satellite Account (TSA)
The *Tourism Satellite Account* (described in the *Tourism Satellite Account: Recommended Methodological Framework 2008*) is, besides the *International Recommendations for Tourism Statistics 2008*, the second international recommendation on tourism statistics that has been developed in a framework of consistency with the System of National Accounts. Both recommendations are mutually consistent and provide the conceptual framework for measuring and analyzing tourism as an economic activity in each country.

As a statistical tool for the economic accounting of tourism, the TSA can be seen as a set of ten summary tables, each with their underlying data and representing a different aspect of the economic data relative to tourism: inbound domestic tourism and outbound tourism expenditure, internal tourism expenditure, production accounts of tourism industries, and the gross value added (GVA) and gross domestic product (GDP) attributable to tourism demand, employment, investment, government consumption, and nonmonetary indicators.

The purpose of a Tourism Satellite Account is to analyze in detail all the aspects of demand for goods and services associated with the activity of visitors, to observe the operational interface with the supply of such goods and services within the economy, and to describe how the supply interacts with other economic activities. It should permit greater internal consistency of tourism statistics with the rest of the statistical system of a country as well as increased international comparability of these data. TSA is highly recommended to evaluate the role that tourism sector performs in the entire economy as well as to allow processing and comparison at an international level.

Primary Data: Surveys
Where the required secondary data support is unavailable for an economic impact analysis, primary data collection from survey sampling is probably the only way forward. This approach can provide useful information where more sophisticated methods are not applicable, but the results should be used with caution due to sampling biases. Careful statistical treatments with the raw data are also needed.

International Tourism: An Overview

In 1994, the United Nations classified three forms of tourism in its *Recommendations on Tourism Statistics*:

- Domestic tourism, involving residents of the given country traveling only within this country
- Inbound tourism, involving nonresidents traveling in the given country
- Outbound tourism, involving residents traveling in another country

International tourist arrivals (overnight visitors) grew by 5% in 2013, reaching a record 1087 million arrivals worldwide, up from 1035 million in 2012.

International tourism receipts are the earnings generated in destination countries from expenditure on accommodation, food and drink, local transport, entertainment, shopping, and other services and goods. In macroeconomic terms, expenditure by international visitors counts as exports for the destination country and as imports for the

country of residency of the visitor. In 2013, international tourism receipts in destinations around the world grew 5% in real terms (taking into account exchange rate fluctuations and inflation) to reach US$ 1159 billion (euro 873 bn). The growth in receipts mirrored the growth in international arrivals (also +5%), confirming the strong correlation between these two key indicators of international tourism.

Table 1 reports the most visited countries in 2013 in terms of the number of international tourist arrivals. Table 2 reports the top ten tourism earner countries for the year 2013.

Tourism, Table 1 International tourist arrivals

Rank	Series[1]	Million		Change (%)	
		2012	2013*	12/11	13*/12
1 France	TF	83.0	..	1.8	..
2 United States	TF	66.7	69.8	6.3	4.7
3 Spain	TF	57.5	60.7	2.3	5.6
4 China	TF	57.7	55.7	0.3	−3.5
5 Italy	TF	46.4	47.7	0.5	2.9
6 Turkey	TF	35.7	37.8	3.0	5.9
7 Germany	TCE	30.4	31.5	7.3	3.7
8 United Kingdom	TF	29.3	31.2	−0.1	6.4
9 Russian Federation	TF	25.7	28.4	13.5	10.2
10 Thailand	TF	22.4	26.5	16.2	18.8

[1]*TF* International tourist arrivals at frontiers (excluding same-day visitors); *TCE* International tourist arrivals at collective tourism establishments.
Source: UNWTO (2014) Tourism highlights

In the ranking by arrivals, Europe leads the growth in absolute terms, while Asia and the Pacific record the fastest relative growth across all the regions. France continues to top the ranking of international tourist arrivals and is third in international tourism receipts. The United States ranks first in receipts and second in arrivals. Spain is still the second largest earner worldwide and the first in Europe and ranks third in arrivals. China moves to fourth in arrivals and remains fourth in receipts. Italy consolidates its fifth place in arrivals and sixth in receipts. Turkey remains sixth in arrivals and 12th in receipts. Thailand enters the top ten arrivals ranking at number ten, climbing amazing five positions, while it moves up two places to seventh in the ranking by tourism receipts. Germany and the United Kingdom remain, respectively, seventh and eighth in arrivals, but move down one place each in terms of earnings to eighth and ninth places, respectively. The Russian Federation completes the top ten ranking by arrivals in ninth place, while the two Chinese Special Administrative Regions Macao and Hong Kong rank, respectively, fifth and tenth in receipts (UNWTO *Tourism Highlights* 2014).

World's Top Tourism Destinations

When ranking the world's top international tourism destinations, it is preferable to take more than a single indicator into account. Ranked according to the two key tourism indicators – international tourist arrivals (Table 1) and international tourism

Tourism, Table 2 International tourist receipts

Rank	US$				Local currencies	
	Billion		Change (%)		Change (%)	
	2012	2013*	12/11	13*/12	12/11	13*/12
1 United States	126.2	139.6	9.2	10.6	9.2	10.6
2 Spain	56.3	60.4	−6.3	7.4	1.5	3.9
3 France	53.6	56.1	−2.2	4.8	6.0	1.3
4 China	50.0	51.7	3.2	3.3	0.8	1.4
5 Macao (China)	43.7	51.6	13.7	18.1	13.2	18.1
6 Italy	41.2	43.9	−4.2	6.6	3.8	3.1
7 Thailand	33.8	42.1	24.4	24.4	26.7	23.1
8 Germany	38.1	41.2	−1.9	8.1	6.3	4.5
9 United Kingdom	36.2	40.6	3.3	12.1	4.8	13.2
10 Hong Kong (China)	33.1	38.9	16.2	17.7	15.8	17.7

Source: UNWTO (2014) Tourism highlights

receipts (Table 2) – it is interesting to note that eight of the top ten destinations appear on both lists, despite showing marked differences in terms of the type of tourists they attract and in average length of stay and spending per trip and per night. In the case of international tourism receipts, changes not only reflect relative performance but also (to a considerable extent) exchange rate fluctuations between national currencies and the US dollar.

The World Tourism Organization also ranks countries on the base of their total expenditure on international tourism for the year 2013. Table 3 reports the top ten countries.

Tourism Distribution Channels

Two of the most important factors influencing the competitiveness and success of a tourist destination are the efficiency and effectiveness of distribution channels (Buhalis and Laws 2001).

Tour operators play a significant role in the promotion and marketing of tourist products. This is particularly true for small and medium accommodation structures and emerging regions which do not have sufficient expertise and financial resources to invest in advertising and in the widespread distribution of their products (Buhalis 2000; Succurro 2008).

The negotiation between tourist firms and intermediaries, primarily aimed at increasing visibility and competitiveness of services supplied, is relevant to the success of both the tourist firms and the tourist destinations as a whole. Moreover, an appropriate promotion and distribution of products have direct consequences on a balanced use of accommodation establishments. High seasonality, indeed, is one of the most problematic aspects in the tourist sector; many destinations suffer from this phenomenon every year.

Both the seasonality problem and the relative importance of traditional tour operators have been strongly affected by the new technologies. Reduced search costs and direct online organization of the trip are two prominent aspects of the tourism industry that have been deeply affected by recent technological advances. Tourism providers, indeed, can now sell their services directly through the web by avoiding intermediaries. Recent studies explore the capacity management issue under time-varying demand (i.e., the seasonality issue), the tourist information acquisition process, and, finally, the impact of online booking on tourism flows and seasonality (Jang 2004; Boffa and Succurro 2012).

ICT and Tourism Distribution

Over the last 20 years, the Internet has changed various facets of social life, creating many social

Tourism, Table 3 Top ten biggest spenders, 2013

	International tourism expenditure (US$ billion)		Local currencies change (%)		Market share (%)	Population (million)	Expenditure per capita (US$)
Rank	2012	2013*	12/11	13*/12	2013*	2013	2013*
1 China	102.0	128.6	37.3	23.8	11.1	1,361	94
2 United States	83.5	86.2	6.7	3.3	7.4	316	273
3 Germany	81.3	85.9	2.5	2.3	7.4	81	1,063
4 Russian Federation	42.8	53.5	36.5	28.9	4.6	143	374
5 United Kingdom	51.3	52.6	2.1	3.5	4.5	64	821
6 France	39.1	42.4	−5.8	4.9	3.7	64	665
7 Canada	35.0	35.2	6.2	3.2	3.0	35	1,002
8 Australia	28.0	28.4	2.1	8.8	2.4	23	1,223
9 Italy	26.4	27.0	−0.3	−1.0	2.3	60	452
10 Brazil	22.2	25.1	4.6	12.9	2.2	198	127

Source: UNWTO (2014) Tourism highlights

concerns (Kim 2010). A large diffusion of the ICT in the tourism sector has improved its social and economic impacts, from which many consumers and organizations can benefit (Minghetti and Buhalis 2010). Indeed, the Internet has grown to be one of the most effective means for tourists to seek information and purchase tourism-related products (Pan and Fesenmaier 2006).

The Internet plays a key role in the development of the tourism industry since it encourages people to travel both by improving access to the destinations and by reducing search costs (Boffa and Succurro 2012). With the advent of e-commerce, indeed, tourism products have become one of the most traded online items.

Technological progress, coupled with regulatory changes, has modified the nature of tourists' search process in at least two directions. First, it has expanded consumption opportunities (e.g., by decreasing the cost of reaching relatively far destination) and, as a result, the expected benefit of searching for a tourist destination. Second, it has decreased the costs of direct search, for example, through the release of faster and more reliable search tools, thanks to the Internet. The two effects contribute to making the direct online search process less time-consuming and more valuable, hence more productive.

Tourism Research

Despite the debate about its definition over the past decades, tourism is commonly recognized as a human activity which can be analyzed from different perspectives. As a field of study, tourism is gradually evolving from a multidisciplinary endeavor into an interdisciplinary stage of research (Tribe and Xiao 2011). Thus, in order to advance the understanding of tourism, it is necessary to integrate economics with other social sciences including law, psychology, sociology, and political science and to seek new holistic approaches and tools.

Numerous disciplinary contributions in diverse areas of research have supported the emergence of tourism as a field of academic study or an autonomous discipline. The epistemology of tourism research, however, is still the subject of ongoing discussion and debate (Benckendorff and Zehrer 2013).

Economic Analysis

In the wider context of tourism knowledge creation, economics has played a significant role.

Tourism generates directly and indirectly an increase in economic activity in the destinations visited, mainly due to demand for goods and services that need to be produced and provided.

In the economic analysis of tourism, one may distinguish between tourism's economic *contribution* which refers to the direct effect of tourism and is measurable through the Tourism Satellite Account (TSA) and tourism's economic *impact* which is a much broader concept encapsulating the direct, indirect, and induced effects of tourism which must be estimated by applying models to provide the simulations (Dwyer et al. 2007; Song et al. 2012). Thus, economic impact studies aim to quantify economic benefits, that is, the net increase in the wealth of residents resulting from tourism, measured in monetary terms, over and above the levels that would prevail in its absence.

Because of the evolution of tourism as an economic activity over the past 50 years, there has been a significant growth of publications in specialized journals (*Annals of Tourism Research*, *Tourism Economics*, *Tourism Management*, *Journal of Travel Research*, to cite some of them) and several key texts on tourism economics (more recently, Dwyer et al. 2010; Stabler et al. 2010; Tribe 2011). Several scientific articles have also attempted to provide an overview of the developments in tourism economics (see, as relatively more recent works, Eadington and Redman 1991; Sinclair 1998; Tremblay 1998; Sinclair et al. 2003; Dwyer et al. 2011.

Macroeconomic Analysis

From a macroeconomic perspective, tourism contributes to both destination competitiveness – defined in different ways and measured by different methodologies – and local, national, and international economic development. Over the last few decades, indeed, a popular topic in tourism research has been the evaluation of the

economic, social, and environmental impact of tourism and its policy implications.

With reference to the economic impact, proponents of the tourism-led growth hypothesis focus on the relationship between tourism and economic growth and emphasize the role of tourism in spurring local investments, exploiting economies of scale, increasing employment, and diffusing technical knowledge (Schubert et al. 2011). The employment effect of tourism, the quality and structure of employment, and the gender wage gap are other well-established and interdisciplinary research areas which also draw on insights from sociology and political science.

With reference to the tourism-led theory, many studies confirm a unidirectional causality from international tourism to real GDP in specific countries or regions, while other studies find evidence of bidirectional relationships. The main criticism faced by the tourism-led growth studies relates to their reliance on the use of a methodology – the Granger causality test – which does not necessarily suggest the real cause-effect relationship (Song et al. 2012).

Note that recent research stresses that tourism does not always increase economic welfare. In fact, a tourism boom may lead to "deindustrialization" in other sectors due to a phenomenon known as the "Dutch Disease effect."

Moreover, as a significant form of international trade flows, tourism also lies within the scope of international economics studies. A number of studies find supportive evidence of the bidirectional causality between international tourism and international trade. Also these studies usually rely on the Granger causality test with the subsequent aforementioned criticism.

With reference to the environmental issues, tourism production and consumption generate environmental consequences, and at the same time, tourism activities are strongly affected by the quality of the environmental resources. In the tourism industry, and differently from the manufacturing industries, the environment is not only an input factor but also a key component of its output (Razumova et al. 2009). For this reason, an increasing attention has been paid to sustainable tourism and climate change topics – both at the micro- and macrolevel of analysis – and to discussions about the appropriate instruments for environmental governance. Several price-based instruments, semi-price instruments such as quotas, and non-price instruments, such as government regulations and industry voluntary management, have been proposed and discussed theoretically in the tourism literature.

Microeconomic Analysis

From a microeconomic perspective, economic studies are complex and cover several topics. In the recent tourism economic literature, departing from the old debate about whether tourism is an industry or a market, it has been commonly recognized that tourism cannot be defined as either an industry or a market. Clarification of this confusion has important implications for economic analysis in this field (Wilson 1998). More specifically, ". . .tourism is a composite product that involves a combination of a variety of goods and services provided by different sectors, such as transport, accommodation, tour operators, travel agencies, visitor attractions, and retailing. Moreover, tourism products are serviced and transacted in different markets" (Song et al. 2012). For this reason, diverse topics have been developed from both the demand and supply perspectives.

Demand Analysis The dominant position of demand analysis and its determinants is still observable in the latest developments in both theoretical and empirical tourism economic studies. Tourism demand is predominantly measured by the number of arrivals, by the level of tourist expenditure (receipts), and, more recently, by the number of tourist nights (length of stay). The key research tasks in tourism demand studies include the selection of the best specified models for modeling and forecasting tourism demand, the identification of the key economic determinants of tourism demand, and the computation of demand elasticities and, associated with globalization, market interdependence.

Supply Analysis The tourism supply analysis usually follows that of industrial economics. The well-known structure-conduct-performance (SCP)

paradigm has provided a useful framework for studying tourism supply from a market perspective. The SCP paradigm suggests that the type of the market structure within which a firm operates (e.g., monopoly, monopolistic competition, oligopoly) determines a firm's conduct (e.g., pricing strategies, marketing investment, innovation) which ultimately affects its overall performance (mainly productivity, efficiency, and long-term growth). A number of empirical studies have tested the SCP paradigm in tourism, but the findings are inconclusive due largely to the different empirical settings and methods used. Newer approaches take into account the dynamic nature of the market – on the base of game theory approach – and its institutional arrangements.

In addition to the above topics, industry agglomeration and clustering, with the economic importance of geographic location, have become an emerging topic in recent tourism supply studies. The new geographical economics, indeed, provides useful perspective for interfirm relationships.

Adjectival Tourism

Many niche or specialty travel forms of tourism have come into a common use by the tourism industry and academics. Among the various forms of tourism: individual tourism, collective tourism, organized tourism, educational tourism, young tourism, third-age tourism, business tourism, sustainable tourism, and ecotourism.

Other forms of "adjectival tourism" include agritourism, birth tourism, culinary tourism, cultural tourism, extreme tourism, geotourism, heritage tourism, medical tourism, nautical tourism, religious tourism, sex tourism, and wildlife tourism.

Sustainable Tourism

Tourism's rapid growth calls for a greater commitment to the principles of sustainability to amplify tourism's benefits and mitigate its possibly negative impacts on the environment and on societies.

The key issues in sustainable tourism, indeed, are defined by the fundamentals of sustainability, external to the literature of tourism research, and linked to science, environment, resource management, global change, human health, economics, and development policy (Buckley 2012).

From a more general perspective, indeed, and as reported by the World Commission on Environment and Development (officially dissolved in December 1987), sustainable development implies "meeting the needs of the present without compromising the ability of future generations to meet their own needs." Thus, sustainable development would at least maintain ecological integrity and diversity to meet human needs.

Tourism researchers turned their attention to social and environmental issues almost four decades ago. Research using the specific term *sustainable tourism* started around two decades ago.

Specifically, "sustainable tourism is envisaged as leading to management of all resources in such a way that economic, social and aesthetic needs can be fulfilled while maintaining cultural integrity, essential ecological processes, biological diversity and life support systems".

Sustainable tourism can be seen as having regard to ecological and sociocultural carrying capacities and includes involving the community of the destination in tourism development planning. It also involves integrating tourism to match current economic and growth policies so as to mitigate some of the negative economic and social impacts of mass tourism.

There is a myriad of definitions for sustainable tourism, including ecotourism, green travel, environmentally and culturally responsible tourism, fair trade, and ethical travel.

Ecotourism

Ecotourism, also known as ecological tourism, is responsible travel to fragile and usually protected areas that strives to be low impact and (often) small scale. It helps educate the traveler, provides funds for conservation, directly benefits the economic development and political empowerment of local communities, and fosters respect for different cultures and for human rights.

Summary/Conclusion/Future Directions

Economics has played a significant role in studying tourism, considered as a key driver of socio-economic progress. In the wider context of tourism knowledge creation, however, there is a need for a holistic approach to tourism analysis. Thus, it would be desirable to integrate economics with other social science disciplines.

References

Benckendorff P, Zehrer A (2013) A network analysis of tourism research. Ann Tour Res 43:121–149

Boffa F, Succurro M (2012) The impact of search cost reduction on seasonality. Ann Tour Res 39:1176–1198

Buckley R (2012) Sustainable tourism: research and reality. Ann Tour Res 39:528–546

Buhalis D (2000) Relationships in the distribution channel of tourism: conflicts between hoteliers and tour operators in the Mediterranean region. Int J Hosp Tour Adm 1:113–139

Buhalis D, Laws E (2001) Tourism distribution channels. Practices, issues and transformations. Continuum, London

Dwyer L, Forsyth P, Spurr R (2007) Contrasting the uses of TSAs and CGE models: measuring tourism yield and productivity. Tour Econ 13:537–551

Dwyer L, Forsyth P, Dwyer W (2010) Tourism economics and policy. Channel View Publications, Bristol

Dwyer L, Forsyth P, Papatheodorou A (2011) Economics of tourism. In: Cooper C (ed) Contemporary tourism reviews. Goodfellow Publishers, Oxford, pp 1–29

Eadington WR, Redman M (1991) Economics and tourism. Ann Tour Res 18:41–56

Jang SS (2004) Mitigating tourism seasonality. A quantitative approach. Ann Tour Res 31:819–836

Kim S (2010) The diffusion of the internet: trend and causes. Soc Sci Res 40:602–613

Manila Declaration on World Tourism. The world tourism conference, Manila, 27 Sept–10 Oct 1980, pp 1–4

Minghetti V, Buhalis D (2010) Digital divide in tourism. J Travel Res 49:267–281

Pan B, Fesenmaier DR (2006) Online information search. Vacation planning process. Ann Tour Res 33:809–832

Razumova M, Lozano J, Rey-Maquieira J (2009) Is environmental regulation harmful for competitiveness? The applicability of the Porter hypothesis to tourism. Tour Anal 14:387–400

Recommendations on Tourism Statistics (1994) Statistical papers, no.83, M. United Nations, New York, p 5. Retrieved 12 July 2010

Schubert SF, Brida JG, Risso WA (2011) The impacts of international tourism demand on economic growth of small economies dependent on tourism. Tour Manage 32:377–385

Sinclair MT (1998) Tourism and economic development: a survey. J Dev Stud 34:1–15

Sinclair MT, Blake A, Sugiyarto G (2003) The economics of tourism. In: Cooper C (ed) Classic reviews in tourism. Channel View Publications, Clevedon, pp 22–54

Song H, Dwyer L, Li G, Cao Z (2012) Tourism economics research: a review and assessment. Ann Tour Res 39:1653–1682

Stabler MJ, Papatheodorou A, Sinclair MT (2010) The economics of tourism, 2nd edn. Routledge, Abingdon

Succurro M (2008) The role of intermediaries in the growth of a lesser developed region: some empirical evidence from Calabria, Italy. Tour Econ 14:393–407

Tremblay P (1998) The economic organization of tourism. Ann Tour Res 25:837–859

Tribe J (2011) The economics of recreation, leisure and tourism. Butterworth-Heinemann, Oxford

Tribe J, Xiao H (2011) Developments in tourism social science. Ann Tour Res 38:7–26

UNWTO (2015) Tourism highlights, 2015 Edition. Madrid, Spain. Available at www.unwto.org/pub

Wilson K (1998) Market/Industry confusion in tourism economic analyses. Ann Tour Res 25:803–817

Tournament Theory

Martin Schneider
Faculty for Economics and Business Administration, University of Paderborn, Paderborn, Germany

Abstract

In tournament theory the effects of competitions in which the best performers are awarded a fixed prize are studied. The tournament idea has been used to explain career patterns in large US law firms and in European judicial hierarchies. It has also been suggested in a prescriptive way as a method to select judges for the US Supreme Court. Tournaments theory helps to understand under which conditions lawyers and judges engage in a rate race to achieve promotion. But important assumptions of the formal tournament models are not met in practice, so real tournaments are unlikely occur in practice. The theory should therefore not be interpreted as an exact descriptive or prescriptive model of behavior but rather as a useful metaphor to help understand empirical patterns.

Synonyms

Contest theory

Fundamentals

Tournament theory utilizes the metaphor of a sport tournament to analyze certain types of competition between players more generally. In a tournament, the prize of a competition is awarded to a fraction of all competitors with the best performance. Hence, a tournament is distinct from other types of competition by two characteristics. First, the winner prize and the fraction of winners are specified in advance of the competition. Second, the prize is awarded not on the basis of some absolute performance standard but on the basis of relative performance (rank-order tournament).

The idea of a tournament has been suggested for the first time within labor economics and human resource management to describe competition among employees for promotion in an organizational hierarchy (Lazear and Rosen 1981; Rosenbaum 1979, 1984). Since then the tournament idea has been most widely applied to explain the pay structure within companies. It has also been adopted in a range of disciplines including law (more on this below), ecology, finance, and psychology (Connelly et al. 2014).

The basic formal model (Lazear and Rosen 1981) is a game with two identical risk-neutral agents and one principal. The principal seeks to elicit effort from the two agents by setting a tournament prize for the winner, i.e., for the agent with the highest output (output can be measured, effort cannot). Each agent maximizes utility by selecting the level of effort, given the impact of effort on the probability of winning, the prize, and the disutility from effort. At the utility-maximizing level of effort, the marginal disutility from effort is equal to the marginal increase in the probability of winning times the prize. Since agents are identical and the probability of winning depends on the other agent's effort, the tournament incites a rat race among the agents. None of the agents has an incentive to reduce effort because effort reduction implies that the other agent will win. The principal can implement the optimal amount of effort by manipulating the prize, i.e., the difference between the winning and the losing pay level.

In terms of incentives effects, this basic model has two main testable hypotheses (Connelly et al. 2014). First, agents' effort increases with the prize, i.e., with the difference between the winning and losing pay level. Second, it is only this prize not the absolute winning pay level that matters for agents' effort.

The basic formal model has been extended in various directions (Connelly et al. 2014). There may be a succession of tournaments at different levels of an organization rather than a single competition. More than two players may compete. Agents may be heterogeneous. They may be able to influence each other's output (e.g., through sabotage). And output may be influenced by luck to varying degrees.

While the basic formal model focuses on deriving a prize that provides an optimal level of incentives, tournaments may also be implemented for other reasons, namely, to select certain players and to commit the principal to a certain policy (pay, promotion) in advance.

Tournaments of Lawyers and Judges

From a law-and-economics perspective, applications of the tournament idea on judiciaries and law firms are particularly interesting. It has been suggested that certain ("elite") law firms in the USA have implemented more or less rigid tournament models in the early twentieth century (Galanter 1994; Galanter and Palay 1990). In essence, these firms have committed to promoting a certain percentage of associates to partners each year, thus incentivizing younger associates and building up pressure in the firm as a whole to maintain revenues. This tournament, it has been argued, can explain the growth of these law firms in the twentieth century. In the big law firms, competition for partnerships is now restricted to the higher ranks ("elastic tournaments") (Galanter and Henderson 2008). Tournament theory and the idea of up-or-out competition have become the main theoretical perspective to

explain the labor market for lawyers in the large law firms in the USA and elsewhere (Wilkins and Gulati 1998).

Tournament theory has also been used to examine competition among judges for assignment or promotion. Judiciaries in civil law countries such as Germany resemble internal labor markets. Judges enter at the bottom of the hierarchy, they are tenured, and they may be promoted to higher judicial positions (Schneider 2004, 2005). The way judges are selected, the fact that pay is attached to judicial positions and observed data on promotions are in line with the idea that a succession of rank-order tournaments takes place between judges. While the tournament model was used here in a descriptive and explanatory way, another application is prescriptive. In order to avoid politicized nomination, it has been suggested that a tournament might be an appropriate method to select judges for the US Supreme Court (Choi and Gulati 2004a). As a possible criterion for promotion, a rank order of judges has been compiled based on various output measures (Choi and Gulati 2004b).

Why Tournaments?

Tournaments may be considered an efficient incentive mechanism. In contrast to an obvious alternative, namely, paying workers directly according to productivity, the measurement of output is easier. This is because, to determine the winner in a tournament, output needs not be measured continuously; an ordinal measure – a rank order – is sufficient. For both lawyers in law firms and for judges, the provision of incentives has been suggested a reason for the practice of tournament-like promotions (Schneider 2005; Choi and Gulati 2004a; Galanter and Palay 1990).

In addition, tournaments may exert favorable selection effects. If productivity in the current job predicts productivity in a job higher up in the hierarchy, then tournaments may help to identify the best lawyers and judges for more important positions in the law firm and the judiciary. This selection effect of a tournament was considered particularly attractive for judicial positions. Here, a tournament promises to be transparent and objective, while alternative mechanisms are plagued with political bickering and tend to bring to office judges based on their political leanings (Choi and Gulati 2004a).

Tournaments, finally, are also a commitment device. Because prize, the rate of winners, and the output measure need to be specified in advance and are observed easily, workers may trust that effort will be rewarded and the employer will not renege on her promise. This argument has been suggested as an important reason for the big law firms to employ tournaments (Galanter and Palay 1990).

Problems of Tournaments

These potentially positive effects of tournaments are diluted by the various problems that tournaments meet (or would meet) in practice. Even in the large private firms, the first application of tournament theory, promotions are governed in many heterogeneous ways that do not fully comply with the tournament model (Gibbs 1994). A similar statement has been made for large US law firms (Wilkins and Gulati 1998).

Many of the assumptions of the tournament idea are not and cannot be fully met. For example, the rate of winners cannot be specified in advance in judicial hierarchies because the vacancies are often determined by the case load, organizational demography, and public budgets. Hence, there is clearly no commitment to a tournament in judicial hierarchies. Similarly, the assumption that competitors do not know each other's performance is hardly met in practice, which in turn reduces the incentive effect.

A number of important problems concern the measurement of output. Competition for promotions can only incentivize and select efficiently if output is measured appropriately. For example, in response to suggesting a tournament of judges for the Supreme Court, it has been argued that output measures referring to the publication and influence of decisions do not fully capture the "virtue" of a good judge (Solum 2004) and are plagued

with inaccuracy (Levy et al. 2010). Similar concerns have been raised with regard to measuring the productivity of associates in law firms (Wilkins and Gulati 1998).

Conclusion

Tournament theory offers an interesting perspective on the competitions between lawyers in large law firms and between judges in judicial hierarchies. The theory predicts that lawyers and judges may sometimes engage in a rat race and that the winners tend to be better than the losers in measured dimensions of productivity. However, the assumptions of a pure rank-order tournament are never fully met in real law firms and real judiciaries. Therefore, tournament theory should not be taken as a descriptive or prescriptive "model" but rather a "metaphor" to help understand empirical patterns (Wilkins and Gulati 1998).

Cross-References

- ▶ German Law System
- ▶ Intrinsic and Extrinsic Motivation

References

Choi S, Gulati M (2004a) A tournament of judges? Calif Law Rev 92:299–322
Choi SJ, Gulati GM (2004b) Choosing the next Supreme Court justice: an empirical ranking of judge performance. South Calif Law Rev 78:23–117
Connelly BL, Tihanyi L, Crook TR, Gangloff KA (2014) Tournament theory: thirty years of contests and competitions. J Manag 40:16–47
Galanter M (1994) Tournament of lawyers: the transformation of the big law firm. University of Chicago Press, Chicago
Galanter M, Henderson W (2008) The elastic tournament: a second transformation of the big law firm. Stanford Law Rev 60:1867–1929
Galanter M, Palay TM (1990) Why the big get bigger: the promotion-to-partner tournament and the growth of large law firms. Va Law Rev 76:747–811
Gibbs M (1994) Testing tournaments? An appraisal of the theory and evidence. Labor Law J 45:493–500
Lazear EP, Rosen S (1981) Rank-order tournaments as optimum labor contracts. J Polit Econ 89:841–864
Levy MK, Stith K, Cabranes JA (2010) The costs of judging judges by the numbers. Yale Law Policy Rev 28:313–323
Rosenbaum JE (1979) Tournament mobility: career patterns in a corporation. Adm Sci Q 24:220–241
Rosenbaum JE (1984) Career mobility in a corporate hierarchy. Academic, New York
Schneider M (2004) Careers in a judicial hierarchy. Int J Manpow 25:431–446
Schneider MR (2005) Judicial career incentives and court performance: an empirical study of the German labour courts of appeal. Eur J Law Econ 20:127–144
Solum LB (2004) A tournament of virtue. Florida State Univ Rev 32:1365–1400
Wilkins DB, Gulati GM (1998) Reconceiving the tournament of lawyers: tracking, seeding, and information control in the internal labor markets of elite law firms. Va Law Rev 84:1581–1681

Traceability

Christophe Charlier
Department of Economics, Université Côte d'Azur CNRS, GREDEG, Nice, France

Definition

"The ability to trace the history, application or location of an entity by means of recorded identifications" in "ISO 8402:1994 Quality management and quality assurance – Vocabulary."

Introduction

A general definition of traceability is found in "ISO 8402:1994 Quality management and quality assurance – Vocabulary" as "The ability to trace the history, application or location of an entity by means of recorded identifications." Traceability is a typical example of a voluntary management practice that predated food safety regulation that ends up entering sanitary public policy. Voluntarily used at the beginning by few operators in high-quality food chains only, traceability has become a mandatory risk management practice for all food operators in Europe. However, legal forms (voluntary vs. mandatory), as well as exigencies of the various traceability systems implemented

around the world, remain very different, even for a same product (e.g., see Schroeder and Tonsor 2012 for bovine traceability).

Traceability is considered from two broad perspectives in economic literature. First, it has been analyzed from firms' standpoint. The determinants of traceability adoption have been explored in this field. The link between traceability adoption and firm characteristics such as their complexity, their hierarchical structure, and the kind of relations they have with downstream suppliers has been put forward (Galliano and Orozco 2011, 2013). The aim of operating under a private standard (that requires traceability) or under a quality label such as a geographical indication has also been presented as an important adoption determinant (Souza Monteiro and Caswell 2009).

Second, traceability has been analyzed from the social standpoint. This part of the literature is driven by the aim of establishing economic rationale for traceability. Whether mandatory or not, traceability is justified by information asymmetry among food business operators along the food chain and among food producers and consumers. This information problem arises because of the "credence good" attribute of food safety or quality. As the nature of this information problem is multidimensional, traceability may be more or less demanding. The economic literature agrees with the idea that traceability may be defined according to three characteristics (Golan et al. 2004): its breadth, defined as the amount of information delivered by the system (i.e., the variety of the items recorded); its depth, characterized by how far back in the supply chain information record is made; and its precision, generated by the tracking unit used. Depending on the "values" chosen for these characteristics, different traceability systems are used.

Whatever its form, traceability cannot in itself insure food safety or food quality. It only gives information. However, by delivering information, traceability performs three functions that impact on food safety (Hobbs 2004) and the way food safety crisis may be managed. First, the possibility of tracking back the origin of food through the supply chain allows for ex post efficient management of food safety crisis (withdrawal of unsafe products from the supply chain) and implies cost reduction as a consequence (health cost, lost productivity, lower product sales, etc.). Second, breaking points along the supply chain and responsible operators are more easily identifiable with traceability, so that liability can be established. Operators are therefore more inclined to comply with regulatory standards. Finally, when coupled with labeling, traceability may convey information to consumers on quality or safety attributes. These links between traceability and food safety may seem obvious. However, in the aftermath of the bovine spongiform encephalopathy crisis, when traceability emerged as a risk management tool, they were debated in the Committee on General Principles of the Codex Alimentarius. Unsurprisingly, traceability has not entered national food safety legislations in the same way. For example, traceability has remained limited to certain products in the USA, whereas it has become mandatory for all food and foodstuff in the EU.

This entry considers the literature dealing with the "social standpoint" on traceability. It first focuses on the European Regulation making traceability mandatory. It then examines voluntary traceability through the two main economic reasons previously underlined: incentives and liability concerns.

Mandatory Versus Voluntary Traceability

The Regulation (EC) No. 178/2002 considers traceability as a risk management tool enabling "accurate and targeted withdrawal" of products from the food chain. The Regulation has a broad scope of application since every stage of the food chain is concerned, demanding that all food products and foodstuffs must be tracked. Breadth and depth of such a system are therefore maximal. However, the Regulation only requires operators to be able to identify "the business from which the food, feed, animal or substance that may be incorporated into food or feed has been supplied." Furthermore, no requirement on batch formation appears. The text implicitly favors a formation of

batches according to the identity to whom the products are sold without taking into account their homogeneity in terms of their inputs' origins. The precision of the mandatory traceability implemented is thus very weak. Charlier and Valceschini (2008) show that with these characteristics, the traceability is constructed "step by step" without requiring a compilation of data that would grow along the food supply chain until the good is sold in the market. The authors show that this traceability system alone may hardly reach the goal assigned by the regulation of a targeted withdrawal of products mainly because of the lack of discipline on the products' batch formation.

Together with traceability, the Regulation (EC) No. 178/2002 stipulates operator's responsibilities on food safety procedures concerning food withdrawal and obligations to inform and cooperate with public authorities in case of a sanitary disruption. Interestingly, these obligations and the proactive behavior required on product withdrawal call for a more stringent traceability than the mandatory one. The implementation of this traceability is therefore implicitly left to operators that have to coordinate their efforts (adopting common rules on the constitution of batches) so that the traceability system adopted by downstream operators does not scramble information produced by traceability efforts of upstream operators (Charlier and Valceschini 2008). This individual choice to implement stricter traceability responds to incentives created by regulatory disposition on operator's responsibilities. Not surprisingly, incentives and liability have been found in the economic literature as the two main drivers for voluntary adoption of traceability.

Incentives and Liability

A significant part of the literature has dealt with traceability through the lens of incentives. In this field, an important characteristic of traceability is that this information system allows allocating the cost of food safety failures to the responsible parties. To do so, traceability is always implemented jointly with another device (e.g., inspection procedures) allowing detecting food safety failures (Starbird and Amanor-Boadu 2006, 2007). Relations among suppliers and buyers of foodstuff are seen as taking place in context of asymmetric information and are generally represented using a principal-agent framework (Resende-Filho and Buhr 2008). The models developed delineate the conditions under which traceability provides incentives to suppliers to deliver safe products. The results depend crucially on various variables such as the food safety failure cost, the production cost of a safe product, the inspection cost, and the cost of allocating failure responsibility. Traceability in this context may be modeled differently. It may be characterized by a "cost allocation factor," i.e., "the proportion of food safety failure costs that can be allocated to the producer responsible for unsafe food" (Starbird and Amanor-Boadu 2007), or by the "probability of finding the source of a problem" (Resende-Filho and Hurley 2012). This cost (or this probability) is an exogenous variable. When traceability is mandatory, it thus depends on policy makers' decisions.

The two main messages of this literature are first that traceability is not unambiguous for food safety since precision in tracking foodstuff and incentive payments in contracts are substitute (Resende-Filho and Hurley 2012) or because incentives to choose contracts selecting safe producers not only depend on the cost allocation factor but also on the importance of the failure costs (Starbird and Amanor-Boadu 2007). Second, depending on the values observed for the various other variables influencing incentives to deliver safe products, policy makers should choose adequate values for traceability.

This ambiguity is also found when incentives come from industry reputation preservation rather than the allocation of the cost of food safety failures. Indeed, when traceability enhances food safety along the food chain, food safety reputation for the industry is created. However, this reputation is a fragile asset and is highly sensible to food safety failures and publicized product recalls. A very localized disruption may disturb the entire food chain if the product withdrawal cannot be targeted. In such a context, traceability allowing targeted and rapid product withdrawal may be

seen as protecting the reputation of the food chain and its profits (Pouliot and Sumners 2013). However, this result on profits presupposes that demand reaction to food safety failures is strong. If this assumption is relaxed, targeted withdrawals of product have the effect of increasing price to which products remaining on the market are sold. This situation benefits to some suppliers thus less inclined to collectively engage in traceability. At an individual level, the expected revenue may be a decreasing function of food safety reached at the industry level (Pouliot and Sumners 2013).

The consideration of liability appears in articles in line with the previous ones dealing with incentives. In a food chain composed of farms, marketers, and consumers, liability incentives are enhanced by traceability and result in higher food safety level (Pouliot and Sumners 2008). Traceability from marketers to farms does not alter marketer's liability but allows the former to impose costs of food safety failures on the latter. The incentives for delivering safe products thus created by traceability increase consumers' willingness to pay for the foodstuff. Indeed, they consume safer food and are more likely compensated in case of a food safety problem. This behavior results in higher incentives to produce safe food for the entire food chain. However, as the industry size (i.e., the number of operators) increases, free-riding behaviors appear. Operators realize that the impact of their investment in food safety on the global food safety is decreasing. As a corollary, the "concentration" of the supply chain would positively impact on food safety. For a given size of a food chain, the larger are the different operators (i.e., the less numerous they are at each step), the less likely is free riding in food safety investment (Pouliot and Sumners 2008).

Discussion

In light of the preceding developments, voluntary traceability clearly raises a coordination issue among the different operators of a same food chain that has the potential to reframe the supply chain. Very few studies have tackled this problem that may constitute an important theme for future research. Banterle and Stranieri (2008) show, for example, that voluntary traceability increases vertical coordination among firms. As both product quality and characteristics of the processes may be tracked (especially within a private standard framework), diverging views among upstream and downstream operators on the kind of traceability that should be implemented through the chain may emerge and have to be resolved. Therefore, the weight of operators, their strategic position within the supply chain, and their capacity to coordinate at a given step of the food chain to face other operators may be seen as central elements for the kind of traceability finally implemented. If food safety objective remains central in this coordination, other economic motives may be at stake and need to be analyzed.

Cross-References

- ▶ Food Safety
- ▶ Geographical Indications
- ▶ Labeling
- ▶ Risk Management, Optimal

References

Banterle A, Stranieri S (2008) The consequences of voluntary traceability system for supply chain relationships. An application of transaction cost economics. Food Policy 33:560–569

Charlier C, Valceschini E (2008) Coordination for traceability in the food chain. A critical appraisal of European regulation. Eur J Law Econ 25:1–15

Galliano D, Orozco L (2011) The determinants of electronic traceability adoption: a firm-level analysis of French agribusiness. Agribusiness 27:379–397

Galliano D, Orozco L (2013) New technologies and firm organization: the case of electronic traceability systems in French agribusiness. Ind Innov 20:22–47

Golan E, Krissoff B, Kuchler F, Calvin L, Nelson K, Price G (2004) Traceability in the US food supply: economic theory and industry studies. Agricultural Economic Report 830, USDA, Economic Research Service

Hobbs JE (2004) Information asymmetry and the role of traceability systems. Agribusiness 20:397–415

Pouliot S, Sumner DA (2008) Traceability, liability, and incentives for food safety and quality. Am J Agric Econ 90:15–27

Pouliot S, Sumner DA (2013) Traceability, product recalls, industry reputation and food safety. Eur Rev Agric Econ 40:121–142
Resende-Filho MA, Buhr BL (2008) A principal-agent model for evaluating the economic value of a traceability system: a case study with injection-site lesion control in fed cattle. Am J Agric Econ 90:1091–1102
Resende-Filho MA, Hurley TM (2012) Information asymmetry and traceability incentives for food safety. Int J Prod Econ 139:596–603
Schroeder TC, Tonsor GT (2012) International cattle ID and traceability: competitive implications for the US. Food Policy 37:31–40
Souza Monteiro DM, Caswell JA (2009) Traceability adoption at the farm level: an empirical analysis of the Portuguese pear industry. Food Policy 34:94–101
Starbird SA, Amanor-Boadu V (2006) Do inspection and traceability provide incentives for food safety? J Agric Resour Econ 31:14–26
Starbird SA, Amanor-Boadu V (2007) Contract selectivity, food safety, and traceability. J Agric Food Ind Organ 5, Article 2

Tradable Emission Rights

► Emissions Trading

Trade Secrets Law

Luigi Alberto Franzoni
University of Bologna, Bologna, Italy

Abstract

Standardization of trade secrets protection was one of the goals of the TRIPs Agreement of 1998. Still, substantial differences across jurisdictions remain. In defining the optimal scope of trade secrets law, lawmakers should consider that strong protection is likely to promote inventiveness but also to retard the diffusion of knowledge and stymie competition.

Definition

Trade secrets law protects firms from unauthorized disclosure of valuable information. The misappropriation of trade secrets generally constitutes an act of unfair competition that triggers civil liability and, possibly, criminal penalties. Standard examples of misappropriation include espionage, breach of nondisclosure agreements, and unauthorized revelation to third parties.

Historically, trade secrets law has its roots in the Middle Ages, in a time when craft guilds jealously protected their specialized knowledge (the "mysteries" of the arts). Within the guild, as well as in master-apprentice relationships, secrecy was the standard, and its violation could be sanctioned with the capital penalty (see Epstein 1998). In modern times, violation of secrecy has been regarded as an act of unfair competition contrasting with the honest practices that should prevail in the business community. Unfair competition was mentioned in the *Paris Convention for the Protection of Industrial Property* of 1883 (art. 10bis), although with no direct reference to the misappropriation of trade secrets. In spite of the convention, huge differences across countries characterize the field of unfair competition law (see De Vrey 2006; Henning-Bodewig 2013).

Commercial and technological information can leak out of firms in many ways:

It can be stolen by employees or third parties (as in the case of information embodied in documents, files, or technological items); it can be obtained by means of subtle espionage techniques (tapping, dumpster diving, etc.); it can be disclosed to third parties by unfaithful employees; it can be memorized and taken away by former employees who start their own business; it can be indirectly devised by competitors by means of reverse engineering; it can be obtained from the scrutiny of documents submitted to regulatory agencies; it can be obtained by rivals by communication with parties related to the information owner (e.g., buyers and suppliers).

The main function of trade secrets law is to clearly tell methods for the acquisition of information that are admitted from those that are not. Those that are not constitute acts of "unfair

competition" and lead to civil and criminal liability. Given the multitude of ways in which commercial and technological information can spill from firms to the others, in most countries trade secrets law provisions are scattered across several branches of the law, including tort law, contract law, intellectual property (IP) law, labor law, and criminal law. On this account, substantial variations exist across legal systems (see European Commission 2013).

At international level, an important definition of trade secrets is provided by the TRIPs Agreement (art. 39.2), which postulates that lawfully acquired business information qualifies as a trade secret only if it (a) is secret, (b) has commercial value because it is secret, and (c) has been subject to reasonable steps to keep it secret. From this definition, we learn that publicly available information and everyday knowledge are not eligible for legal protection; valueless information and information not subject to reasonable protection do not qualify as trade secrets. All countries belonging to the WTO should make sure that legal protection is granted to trade secrets against acts of misappropriation which include "breach of contract, breach of confidence, and inducement to breach" (TRIPS, footnote 10). The task to define the precise set of activities falling under the "misappropriation" category lies with individual countries, which might be more or less strict on this account. In turn, misappropriation leads to remedies that usually include injunctive relief and damage awards. The latter are typically commensurate with the actual loss to the trade secret owner or the unjust enrichment of the party that has misappropriated the secret. In most countries, courts can also set a reasonable royalty for the use of the secret (for a limited time span).

From a policy perspective, the main issue raised by trade secrets law is the optimal scope of the protection granted to the owners of undisclosed information (which conducts are forbidden and which are not). Some polar conducts can easily be categorized: the theft of documents is undoubtedly unlawful, while reverse engineering tends to be lawful everywhere. With respect to other conducts, courts and lawmakers offer a variety of positions. For example, courts and lawmakers can be more or less lenient with respect to key employees who leave their company to work for a competitor. In some jurisdictions, this conduct can lead to an unfair competition suit, under the doctrine of inevitable disclosure. In some other jurisdictions, e.g., in California, workers' mobility finds little impediments in trade secrets law (see Gilson 1999). In deciding the strength of trade secrets protection, lawmakers need to keep in mind that strong protection comes at the cost of reduced labor mobility and lower diffusion of technological knowledge (see Fosfuri and Rønde 2004).

More generally, trade secrets law has been credited with the following beneficial effect (see Lemley 2011). First, by preventing unwarranted diffusion of valuable information, trade secrets law provides a competitive edge to the original producer of the information. With respect to innovative knowledge, for instance, the head start advantage provided by secrecy is regarded as important by most companies (see Cohen et al. 2000). Second, trade secrets law allows firms to reduce the self-protection expenditure: thanks to the legal obstacles against unwanted disclosure, firms can more freely organize their units and share information across their members (see Rønde 2001). In the absence of legal protection, costly measures would have to be taken to reduce the probability of leakage. In fact, reduction of self-protection expenditure is the main function that Landes and Posner (2003) credit to IP law. Facilitated information sharing can also concern contracting parties outside the firm. Effective enforcement of nondisclosure agreements facilitates the transmission and sale of information from the producer to third parties. In this sense, nondisclosure agreements represent a partial solution to Arrow's information paradox (Arrow 1962), which postulates the unavailability of restitutory remedies for unwarranted information disclosure (once shared, information cannot be returned).

At the same time, trade secrets law also produces social costs. First, by limiting the circulation of information, trade secrets law may retard imitation and stymie technological progress. It has been noted, in fact, that strong technological

spillover characterizes some fast-evolving technological districts like the Silicon Valley, where high levels of labor mobility speed up the diffusion of innovative knowledge (Saxenian 1996). Information sharing facilitates the expansion of the stock of public knowledge, giving birth to new forms of collective invention (Allen 1983; von Hippel and von Krogh 2011).

A further cost brought about by trade secrets law concerns the relationship between secrecy and patent protection (explored, more generally, by Hall et al. 2014). If trade secrets law is strong, inventors have weaker incentives to rely on the patent system. Hence, fewer inventions are disclosed in the patent applications and the stock of public knowledge may advance less rapidly. Here, the issue is whether patents or trade secrecy are better protection tools from a social point of view. Patents have limited duration (normally 20 years) and require disclosure of the invention. Trade secrets can potentially last forever and, by definition, are not disclosed. This implies that either nobody has access to the information – and the original owner retains market power forever – or that third parties have to waste resources to rediscover that original invention (for the purpose to market it directly or to improve upon it). The comparison between the two forms of protection hinges on the nature of the innovation process (how many firms have the capacity to come up with the original invention, the extent of the research spillovers) and the nature of competition across firms upon duplication (under trade secrecy) (see Denicolò and Franzoni 2012).

The impact of trade secrets law on the propensity to patent, however, should not be overestimated. In fact, the subject matter of trade secrets law is infinitely broader than that of patent law. This is because nearly any type of commercial and technological information is eligible for trade secrets protection, while only inventions that meet the originality and non-obviousness requirements qualify for patent protection. A recent investigation by Hall et al. (2013) reveals that only 5% of innovative UK firms rely on the patent system. All of them, one way or another, rely on secrecy.

Cross-References

- ▶ Competitive Neutrality
- ▶ Innovation
- ▶ Intellectual Property: Economic Justification

References

Allen RC (1983) Collective invention. J Econ Behav Organ 4(1):1–24

Arrow K (1962) Economic welfare and the allocation of resources for invention. In: The rate and direction of inventive activity: economic and social factors. Princeton University Press, Princeton

Cohen WM, Nelson RR, Walsh JP (2000) Protecting their intellectual assets: appropriability conditions and why US manufacturing firms patent (or not) (No. w7552), National Bureau of Economic Research

De Vrey R (2006) Towards a European unfair competition law: a clash between legal families. Martinus Nijhoff Publishers, Leiden

Denicolò V, Franzoni LA (2012) Weak intellectual property rights, research spillovers, and the incentive to innovate. Am Law Econ Rev 14(1):111–140

Epstein SR (1998) Craft guilds, apprenticeship, and technological change in preindustrial Europe. J Econ Hist 58(03):684–713

European Commission (2013) Study on trade secrets and confidential business information in the internal market. Downloadable at http://ec.europa.eu/internal_market/iprenforcement/trade_secrets/

Fosfuri A, Rønde T (2004) High-tech clusters, technology spillovers, and trade secret laws. Int J Ind Organ 22(1):45–65

Gilson RJ (1999) Legal infrastructure of high technology industrial districts: Silicon Valley, route 128, and covenants not to compete. NY Univ Law Rev 74(3): 575–629

Hall B, Helmers C, Rogers M, Sena V (2013) The importance (or not) of patents to UK firms. Oxf Econ Pap 65(3):603–629

Hall B, Helmers C, Rogers M, Sena V (2014) The choice between formal and informal intellectual property: a review. J Econ Lit 52(2):375–423

Henning-Bodewig F (ed) (2013) International handbook of unfair competition. Beck, Hart, Oxford and Nomos, Baden Baden

Landes W, Posner R (2003) The economic structure of intellectual property law. Harvard Universtiy Press, Harvard

Lemley MA (2011) The surprising virtues of treating trade secrets as IP rights. In: Dreyfuss R, Strandburg K (eds) The law and theory of trade secrecy: a handbook of contemporary research. Edward Elgar Publishing, Cheltenham

Rønde T (2001) Trade secrets and information sharing. J Econ Manag Strateg 10(3):391–417

Saxenian A (1996) Regional advantage: culture and competition in Silicon Valley and route 128. Harvard University Press, Harvard

Von Hippel E, von Krogh G (2011) Open innovation and the private-collective model for innovation incentives. In: Dreyfuss R, Strandburg K (eds) The law and theory of trade secrecy: a handbook of contemporary research. Edward Elgar Publishing, Cheltenham

Tradeable Discharge Permits

► Transferable Discharge Permits

Trademark Dilution

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

Definition

A trademark dilution claim is raised whenever a new trademark, albeit it does not confuse consumers, produces detriment to the distinctive character of another trademark. Then the use of new trademark, although used in a noncompeting market, can still be forbidden because it weakens the distinctiveness of the already existing trademark.

Main Characteristics

Trademark dilution is a special kind of infringement that does not involve a direct violation, but it is rather a theoretically legitimate behavior that can nevertheless compromise the distinguishing effect of a given trademark (Economides 1998). It is forbidden by recently amended laws, and it is generally raised in the case of so-called famous or strong trademarks. Antidilution measures essentially enshrine the accomplished transformation of trademark into a property right over a sign and its semantic sphere (Beebe 2004; Ramello 2006). More precisely the goal of antidilution clauses is no longer to protect the informational value of the trademark and avoid consumer confusion, for which the existing regulations were sufficient, but rather to punish behaviors that may indirectly encroach upon this newly forged semantic sphere and its economic exploitation (Landes and Posner 2003). The dilution claim is generally raised whenever a trademark – even if it does not confuse consumers – produces "detriment to the distinctive character" of another famous mark (Adidas-Salomon AG v. Fitnessworld Trading Ltd., 2003, 1 C.M.L.R. 14) according to European law, or results in "lessening of the capacity of a famous mark to identify goods and services" according to US law (Lanham Act, Section 43c, 15 USC Section 112). The main idea is that the use of a sign somehow linkable to a famous one in a noncompeting market can still be forbidden because it weakens the distinctiveness (and most importantly the differentiation effect) of the famous mark. This can happen by "tarnishment" if applied to inferior-quality goods that might lower consumers' opinion of the famous mark, or by "blurring" when there is a sort of semantic free-riding, so that the new sign does not confuse the consumer but still indirectly exploits, and thereby impoverishes, the distinctiveness of the famous mark (Lunney 1999).

The general legislative principle that emerges is thus to preserve the semantic capital of a given trademark, in the market and the minds of consumers, by the difficult route of defining property boundaries in the semantic sphere (Ramello and Silva 2006). Such a solution requires directly protecting the signified through a new intellectual property right that extends beyond the tangible market to cover the market for meanings (Ramello 2013). By so doing, the law protects firms' investments in creating new semiotic products and confirms the altered function of the trademark, which now assigns a legal and economic monopoly over a broad relation between signifier and signified (Ramello and Silva 2006).

The enactment of antidilution measures, after being pushed back several times, was ultimately achieved through protracted lobbying efforts although some scholars expressively speak of "doctrinal puzzlement" (McCarthy 2004) and of "the death of common sense" (Lemley 1999). However, besides these criticisms, the antidilutions measures mark the metamorphosis of trademark, born as an informational tool originally intended as an adjunct to the market for safeguarding consumers, into a property right over a sign and its semantic sphere (McClure 1996). In the meanwhile, it marks the "divorce" of trademarks from the goods they are supposed to represent, which in fact heralds their own "commodification" (Lemley 1999).

Cross-References

- ▶ Intellectual Property: Economic Justification
- ▶ Trademarks and the Economic Dimensions of Trademark Law in Europe and Beyond

References

Beebe B (2004) The semiotics of trademark law. UCLA Law Rev 51:621–704

Economides N (1998) Trademarks. In: Newman P (ed) The new Palgrave dictionary of economics and the law. Macmillan, London, pp 601–603

Landes WM, Posner RA (2003) The economic structure of intellectual property law. Harvard University Press, Cambridge, MA

Lemley MA (1999) The modern Lanham act and the death of common sense. Yale Law J 108:1687–1715

Lunney GS Jr (1999) Trademark monopolies. Emory Law J 48:367–487

McCarthy JT (2004) Dilution of a trademark: European and United States law compared. Trademark Rep 94:1163–1181

McClure DM (1996) Trademarks and competition: the recent history. Law Contemp Probl 59:13–43

Ramello GB (2006) What's in a sign? Trademark law and economic theory. J Econ Surv 20:547–565

Ramello GB (2013) The multi-layered action of trademark: meaning, law and market. In: Miceli TJ, Baker MJ (eds) Research handbook on economic models of law. Edward Elgar, Cheltenham

Ramello GB, Silva F (2006) Appropriating signs and meaning: the elusive economics of trademark. Ind Corp Chang 15:937–963

Trademarks and the Economic Dimensions of Trademark Law in Europe and Beyond

P. Sean Morris
Faculty of Law, University of Helsinki, Helsinki, Finland

Abstract

The economic analysis of trademark law continues to draw a number of commentaries, yet more and more, the courts are not factoring concrete economic analysis of trademark law and trademark protection in their decisions. In this entry I give an overview and status of trademarks from a law and economic perspective and suggest that trademark laws need to respond to the economic dimension that occurs on the market and consumer economic behaviour.

Introduction

Trademarks are signs that communicate the economic interest of goods manufacturers or service providers to customers with valuable information to influence their purchasing behavior. Signs which constitutes a trademark are capable of graphical representation, in particular words, letters, numbers, shape of goods, or their packaging, in as much, such signs can distinguish the goods and services of different providers. Thus, one of the underlying functions of a trademark is to provide information about source and origin of goods and services. A trademark is granted and regulated under the trademark laws applicable to the territory, national or *supra-federal*, for which it is applied, and examples of supra-federal trademark legislations are the *Trade Mark Directive* (TMD 2008) in Europe or the *Community Trade Mark Regulation* (CTMR 2009) both of which are for trademark regulation and harmonization in the European Union (EU).

A Federal trademark legislation is, for example, the *Lanham Act* in the United States America

(USA), while a national trademark legislation is, for example, the *Trade Marks Act 1994* in the United Kingdom (UK) or the Swedish Trademark Act (2010: 1877). These acts together constitute the legal regulation for trademarks in their respective territories, although the EU's *supra-federal* TMD and national trademark laws coexist for the harmonization of trademark law throughout the EU's internal market. In the USA, the passage of the *Lanham Act* in 1946 saw the US trademark law as "harmonized" so to speak. The situation in the EU is somewhat different given that the road to harmonization of European trademark law has only been a recent phenomenon with the first TMD of 1989 and revisions in 2008 and 2014. A new trademark directive and trademark regulation in the European Union came into force in January and March 2016 respectively.

Trademarks in Law and Economics

In cases such as *Intel v CPM*, the European Union Court of Justice (EUCJ) introduced the notion of consumer *economic behavior* in the context of European trademark jurisprudence. The recognition by the Court of a change in the economic behavior of consumers was arguably an overt acknowledgement of the role of economic analysis of trademarks. Trademarks as a branch of intellectual property rights have always been the gatekeeper of how businesses function in a competitive economy, and, as the face of goods, products, and services, trademarks' economic functions and regulation are essential to both the rights holder and how competitors enter the market.

Because trademarks are signs that represent the economic identity of goods and services and, in this regard, they serves as the bridge that links customers' economic behaviors to the proprietor of goods and services, then trademarks are arguably the most important aspect of intellectual property rights in a market economy because unlike other regimes such as patents and copyrights, a trademark, albeit with renewals, is essentially perpetual and generally outlives the initial rights holder(s). This perpetual nature of trademarks demonstrates that they are important in how goods, products, and services are placed on the market to guarantee fair trade among the many players that operate in a dynamic market. A dynamic market will always allow flexibility in trademark use through licensing or other methods for trademark owners to encourage fair competition.

But another dynamism of the market is the economic behavior of trademarks, and perhaps it was no surprise that the EUCJ in *Intel v CPM* overtly calls for the economic analysis of trademarks when trademark infringement claims are made. That decision concerned changes in the economic behavior of the average European consumer and how such changes could cause damage to a reputable trademark. This concern by the EUCJ highlights the need for more understanding of the economic dimensions of trademark law and the *real* functions of trademarks in a market economy and its impact on the economic welfare of consumers.

Although the courts sporadically address the law and economic approaches to trademarks, there has long been a growing body of literature that discusses trademarks in a law and economic context. Some of the early scholars that applied law and economics approaches to trademarks include Chamberlin (1933) and Robinson (1933) when their theories of imperfect competition are factored in and, in particular, Chamberlin's reference to "product differentiation," while other scholars (Papandreou 1956; Brown 1948; Mueller 1968) also offered their perspectives on the economics of trademarks. But some of the major works on the economic analysis of trademark law emerged in the late 1980s with those by Landes and Posner (1987) and Economides (1988) and then other contributions by scholars such as Lunney (1999), Ramello (2006), Barnes (2006), and Griffiths (2011).

These approaches to the economic dimension of trademark law incorporate the economic analysis of law. Arguably, the study of law and economics of trademark law began with the great transformation in the study of law and economics as a whole that had roots in the USA at the turn of the twentieth century. Legal luminaries such as

Holmes (1897) helped to sound the trumpet for the law and economic approach when he observed that "the man of statistics and the master of economics" were ideally good tools for the law.

The Chicago School has long been seen as the driving force behind the study of law and economics in the twentieth century with renowned economists in the 1940s and 1950s that gradually built upon *Marshallian* (1890) price theory or other *canons* such as Knight (1933). But the true form of law and economics only took off in the 1960s where economists such as Stigler (1961), Director (1964), and Coase (1960) oversaw the ascent of Chicago approaches to law and economics. Stigler (1961) played an essential role in the development of law and economics of trademarks when he developed a model of optimal consumer search behavior under which advertising reduces consumers' search costs.

There are critics of the 1960s development of law and economics under the auspices of the Chicago School, and some have even argued that the study of law and economics is rather old given that "the study of the interrelations between legal and economic progress is as old as economics itself" (Medema 2010, p. 160). But even if the origins of law and economics can be disputed, there is without a doubt that the modern phenomena that we understand today as law and economics are essentially a phenomenon that gradually developed as the Chicago School of thought in the USA blossomed in the 1960s and 1970s, even to the point where "lemons" (problems with post-sale used cars) were used to analogize information symmetry (Akerlof 1970) or, in trademark terms, the rights holder have a better idea of the quality his mark represents.

By the 1970s, other Chicagoans such as Posner had also emerged fully in his own right applying economic analyses of the law (Posner 1973), and in 1987 he produced with a colleague a seminal article that reflects similar arguments raised in Stigler's information advertising paper (1961). The economics of advertising, as per Stigler and the search theory, essentially served as the launch pad for economic analyses of trademarks and arguably because of information that advertising portrays in particular communicating the trademark to the consumer, even if contemporary scholars believe that "advertising is patently uninformative" (Jordan and Rubin 2008, p. 19). As a result of the broader study of law and economics, the Chicago approach was transposed to a number of other areas such as tax law, antitrust law, criminal law, contract law, and tort law, among others (Cooter and Ulen 2012; Miceli 1997). The law and economics approach gradually moved on to the study of intellectual property rights which covered mainly patents, copyrights, and much later trademarks, and in a 1987 joint paper by Landes and Posner, the law and economics approach was applied to trademark law and to revolutionize the study of trademarks and trademark law in an economic context.

Landes and Posner (1987) in their seminal article which nowadays stands as the cornerstone on the economic analysis of trademark law have argued that a trademark's essential function from an economic perspective is to reduce consumer search costs. Their arguments are steeped in the Chicago School of economic tradition, in particular, what I think was a reflection of Stigler's 1961 paper on the economics of information.

An earlier paper dealing with the economics of information, albeit from a legal perspective, also discussed advertising, which broadly encompasses trademarks (Brown 1948); however, this was prior to full development of the Chicago School. Economides (1988) who offered a purely positive economic theory on the economic analysis of trademarks noted in particular that trademarks can serve as a barrier to entry since the trademark "can also serve to increase welfare through the reduction of an excess number of brands" (Economides, p. 536). This approach, and like many others, that offers an economic perspective falls within microeconomic theory that concerns how individuals make rational decisions (behavioral theory), and it is largely responsible for economic analysis of law and by extent trademark law. The approach in the literature that covers the economic analysis of trademark law has essentially still kept at the forefront the problems of trademarks in the competitive economic space of society in their dual role to promote competition and to reduce consumer search costs.

Trademark Monopolies and Economic Effects

In regulatory instruments such as the CTMR, it is explicitly noted in its recitals the desirability to promote "harmonious development of economic activities" using trademarks. In this context, there is an economic rationale for trademark law since quality and origin suggest an economic incentive for the trademark owner while giving assurances to consumers. In this regard, trademark law is the principal arbiter that protects consumers in their economic decision-making when shopping for goods that are protected under trademark law. Yet, it is those very same trademark laws that also give the trademark owner a monopoly lease to use a specific mark to designate his goods or services. The guarantees that the law provides for trademark owners include a right to prevent others from using the mark and also a *right* to enhance the creative and innovative process in a broad sense.

The argument that trademarks are monopolies is not entirely new and has been around for ages, for example, in the English case *Blanchard v Hill* (1742), trademarks were identified as "one of those monopolies." What this means, even in contemporary times, is that the exclusivity and absoluteness of trademarks make them one of those monopolies that are subjected to being abused in an anticompetitive way and affect how consumer spending decisions are made. As monopolies, the welfare effect on consumers are negative since their spending powers are dictated by the trademark that they are locked onto, which in turn creates monopoly profits for the trademark owner who has the ability to influence the purchasing habit of the consumer. Furthermore, as a branch of intellectual property rights, trademarks are generally seen through the lens of state-granted monopolies, and the literature that emerged over the last six decades or so generally paints trademarks as monopolies (Lunney 1999).

In a number of ways, trademarks create problems on both the market and for the consumer. For consumers the problem of trademark monopolies is that, contrary to reducing search costs, trademarks lock consumers in *perpetuity* – whereas consumers are loyal only to the manufacturers of the goods the trademark represents. Arguably then, consumer search costs are then limited since they do not engage in the option of the alternative and or at times (naively) assume the mark to which they are loyal represents quality. However, on occasions, consumers may find alternatives that allow them to be better off. For the market, the monopoly in trademarks erects barriers to entry for new goods on the market. This barrier to entry, in particular contemporary times, can arise from an abundance of registered trademarks that are not in use or competitors unwilling to license a mark to new entrant that will compete with them on the market. These observations highlight the gray zone in which trademarks must operate catering to the consumer and the rights holder.

However, the legal reality of trademarks' economic effects is perhaps those that are articulated by the Courts since the Courts themselves help to shape the economic activity of markets. In *Qualitex Co. v Jacobson*, a US court was perhaps eager to formally recognize the economic dimension of trademarks and the law when it noted that consumers were essentially rational decision-makers because trademarks allow them to make better purchasing decisions and reduce their search costs (*Qualitex Co. v Jacobson*, 163–164). In Europe, where the harmonization of trademark law is an integral part of the EU's internal market, the EUCJ has often echoed the economic dimension of trademarks, such as in *Intel v CPM* where it observed the possibility of change in economic behavior due to trademark use.

Perhaps the EUCJ's more broad reading of the economic dimension of trademark protection and use was in *L'Oreal v Bellure* where the Court spoke of trademark's investment function. This (additional) interpretation of trademark functions suggests the wide economic value of trademarks and the role trademarks play in a fully integrated and free market. The value of a company's trademark is often the most important piece of asset that a company holds and can leverage with either for investments or loans in the company. The *L'Oreal v Bellure* Court was indeed right when it pointed out the investment function of trademarks

because the value in a trademark is built over time due to the actual investments in a trademark such as advertising or the reputation that the mark earned. As a valuable asset within a company's portfolio trademarks plays an integral investment function (even if *L'Oreal v Bellure*) did not have this reasoning in mind. The implication here is that the economic function of a trademark goes beyond quality and source function but to that of economic value represented by, a form of *inter*- and *intra*-investments which are acquired over time, thereby creating a market for "goodwill."

But despite the inherent monopoly in trademarks, it is the consumer that benefits the most from trademarking activities since the investment function that trademark performs contains a positive spillover when there is increase in trademarking activities. This spillover is the promotion of more competition and the ability of consumers to rationally select goods that are based on their purchasing power. The trademark would normally signal this and influences how the consumer behaves. In dynamic markets such as the EU where national trademarks operate alongside *European* trademarks, consumers respond differently to *external* trademarks that are not well known or where such trademark does not send quality signal. Therefore, consumers can sometimes be suspicious of goods with unknown reputation that originates in a different country. Another factor that also influences consumer economic behavior toward internal and external trademarks in the EU is that of national consciousness and pricing. Even where a trademark signals quality in a reputable good and a new entrant uses a similar trademark to compete, the Courts have cautioned that only a change in consumer's economic behavior could harm the reputable mark. In *Intel v CPM*, the EUCJ introduced the concept of economic behavior in the trademark *lingua*, suggesting the wider role of economic analyses of trademark law and behavioral patterns of consumers:

> the use of the later mark is or would be detrimental to the distinctive character of the earlier mark requires evidence of a change in the economic behaviour of the average consumer of the goods or services for which the earlier mark was registered and consequent on the use of the later mark, or a serious likelihood that such a change will occur in the future. (Intel v CPM, para. 77)

From an economic point of view, the Court cautiously warns that economic changes can affect trademark use and protection. The reasoning by the EUCJ that only a change in the economic behavior of the average consumer could threaten Intel's – INTEL mark if those consumers flock to CPM's – INTELMARK trademark broadens the scope for economic analyses of trademark use and protection and how economic evidence is assessed within the confines of the law. Intel did not provide any (economic) evidence that CPM's INTELMARK was causing harm to its economic activities, and therefore the Court found no reason that Intel's mark was being threatened.

Perhaps one of the wider implications of the *Intel v CPM* ruling by the EUCJ and its notion of change in economic behavior in the average consumer is to look to the broader economic theory of public goods and how that is related to trademark law. Given that Intel's mark is a reputable mark, there is no doubt that it creates positive externalities and therefore a magnet for competitors to free ride on. Trademarks are in a sense - public goods - and therefore an opportunity for competitors to free ride on the good will of reputable marks. In other words, information is a public good, even when embedded in a trademark. Barnes (2006) has similarly argued that trademark owners are information creators and such information is widely available for "all people to use" and hence, arguably is a public good. For trademark law, the implication is even grave, given that, as Barnes (2007) observed, both dynamic allocative efficiency and static allocative efficiency may emerge.

Because trademark protection in Europe and globally is expanding, there is an urgent need to factor in empirical evidence based on the economic behavioral patterns of consumers in trademark litigations that can be used to help the Courts arrive at conclusions that are based on economic theories. Such empirical evidence should go beyond the mere surveys that are often used in

trademark litigation given that excessive trademarks are "goods in their own right" (*Plasticolor v Ford*, 1989, 1332). It is also useful, when discussing economic approaches of trademark law to consider other areas competition law. This is because empirical evidence suggests that trademarks are anticompetitive, for example, by serving as tying products (Smirti 1976) or due to their inherent exclusivity or market power (Morris 2012, 2013), and whether such exclusivity is being abused by trademark owners.

The market economies in which trademarks exist are expanding, and the movement of global commerce has put into sharp focus trademark use and protection and the scope of trademark law. In trademark legislations, whether international instruments such as the Trade-Related Aspects of Intellectual Property Rights (TRIPS), regional harmonization efforts at the EU, and national trademark laws, there is the need to factor in "economic effects" of those laws and how they relate to other areas of regulation in the free market.

Trademarks are no longer seen as *ought* to be protected but rather how to assess the economic impact of trademarks that are currently protected and are in use and what are the economic causes of trademark nonuse. The economic impact of trademarks that are in use drives consumer economic behavior and how the markets respond to any perceivable changes in economic behavior, and this warrants fresh approaches to the economic dimension of trademark law.

Such economic dimension should move into a new direction beyond the predominant function (s) of trademark theory such as the search cost theory or the origin source theory. This new direction may take on, for example, the reduction of barriers to entry by allowing greater flexibility in the use of trademarks by competitors, or how the economic impact of product differentiation in trademarks affects the normative process in which trademarks must operate.

Ultimately, it is the trademark laws that can respond to the normative process, and perhaps, similar to the EUCJ's recognition of change of economic behavior by the average consumer (also echoed in *Environmental Manufacturing v OHIM*, 2013), trademark laws, both in Europe and beyond, can begin to reform by responding to changes in the market economy which ultimately leads to greater efficiency in the competitive process for the creation of wealth. Nevertheless, the change of economic behavior test that the European Courts advocate holds interesting interpretation for the economic analyses of trademark law both in Europe and beyond, since this new direction in trademark law interpretation has raised the threshold for trademark infringement analysis.

Cross-References

- Behavioral Law and Economics
- Economic Analysis of Law
- Efficiency
- Innovation
- Law and Economics
- Trademark Dilution

References

Akerlof G (1970) The market for lemons: quality uncertainty and the market for mechanisms. Q J Econ 84:488–500

Barnes DW (2006) A new economics of trademarks. Northwest J Technol Intellect Prop Law 5:22–67

Barnes DW (2007) Trademark externalities. Yale J Law Technol 10:1–44

Brown R Jr (1948) Advertising and the public interest: legal protection of trade symbols. Yale Law J 57:1165–1206

Chamberlin E (1933) The theory of monopolistic competition: a re-orientation of the theory of value. Harvard University Press, Cambridge

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

Cooter C, Ulen T (2012) Law and economics, 6th edn. Pearson, Boston

Director A (1964) The parity of the economic market place. J Law Econ 7:1–10

Economides N (1988) The economics of trademarks. Trademark Rep 78:523–539

Griffiths A (2011) An economic perspective on trade mark law. Edward Elgar, Cheltenham

Holmes O (1987) The path of law. Harv Law Rev 10:457–478

Jordan E, Rubin P (2008) An economic analysis of the law of false advertising. In: Mialon HM, Rubin PH (eds) Economics, law and individual rights. Routledge, London, pp 18–43

Knight F (1933) The economics of organization. University of Chicago, Chicago
Landes W, Posner R (1987) Trademark law: an economic perspective. J Law Econ 30:265–309
Lunney G Jr (1999) Trademark monopolies. Emory Law J 48:367–487
Marshall A (1890) Principles of economics. Macmillan, London
Medema S (2010) Chicago law and economics. In: Ross E (ed) The Elgar companion to the Chicago school of economics. Edward Elgar, Cheltenham, pp 160–174
Miceli T (1997) Economics of the law: torts, contracts, property, litigation. Oxford University Press, New York
Morris PS (2012) The economics of distinctiveness: the road to monopolization in trademark law. Loyola Int Comp Law Rev 33:321–386
Morris PS (2013) Trademarks as sources of market power. Working Paper
Mueller C (1968) Sources of monopoly power: a phenomenon called 'Product Differentiation'. Am Univ Law Rev 18:1–42
Papandreou A (1956) The economic effect of trademarks. Calif Law Rev 44:503–510
Posner R (1973) Economic analysis of law, 1st edn. University of Chicago Press, Chicago
Ramello G (2006) What is a sign? Trademark law and economic theory. J Econ Surv 20:547–565
Robinson J (1933) The economics of imperfect information. Macmillan, London
Smirti S (1976) Trademarks as tying products: the presumption of economic power. St Johns Law Rev 50:689–724
Stigler G (1961) The economics of information. J Pol Econ 69:213–225

Cases and Legislations

Blanchard v Hill, 26 Eng Rep 692 [1742]
Case C-252/07, Intel Corporation Inc v CPM United Kingdom Ltd [2008], ECR I – 8823
Case C-383/12 P, Environmental Manufacturing LLP v OHIM [2013], ECR I – 0000
Case C-487/07, L'Oreal SA v Bellure NV [2009] ECR I – 05185
Council Regulation (EC) No 207/2009 of 26 February 2009 on the Community Trade Mark, OJ L 78/1, 24.3.2009 (CTMR)
Directive 2008/95/EC of the European Parliament and of the Council of 22 October 2008 to approximate the laws of the Member States Relating to Trade Marks, OJ L 299/25, 8.11.2008 (TMD)
Joined Cases C-236 to C-238/08, Google France v Louis Vuitton Malletier [2010], ECR I – 02417
Lanham (Trademark) Act, 15 U.S.C. s. 1051, 60 Stat 427 [1946]
Plasticolor Molded Products v Ford Motor Co., 713 F. Supp 1329 [1989]
Qualitex Co. v Jacobson Products Co., 514 US S. Ct. 159 [1995]
Trade Marks Act (UK) [1994]
Trademark Act (Sweden) [2010]

Trading of Allowances

▶ Transferable Discharge Permits

Traffic Light Violation

▶ Traffic Lights Violations

Traffic Lights Violations

Laurent Carnis
University Paris Est and IFSTTAR, Noisy-le-Grand, France

Abstract

Traffic light running is a violation of Highway Code that endangers the offender as well as other road users. The consequences of crash risk induced by this reckless behavior can be interpreted as a social cost and call for public intervention. Several tools are available for policy makers to enforce traffic light obedience. However, the cost of public intervention must be in line with the social cost of violations.

Although there is a sizable literature dealing with traffic light running, researchers generally focus on the predictors of such behavior and the impact of the countermeasures. This entry presents a literature overview from an economic perspective and proposes an economic analysis of traffic light violations and their regulation.

Synonyms

Traffic light violation

The author thanks E. Kemel for comments and suggestions on a previous version. The usual caveat applies.

Introduction

Red light running (RLR) is a violation of traffic rules that endangers the offender as well as other road users. A red light isolation is established when a driver fails to stop at a red signal indication. The crash risk induced by this reckless behavior calls for public intervention. Indeed, RLR can be considered as an external effect imposed upon road crash victims. A social cost is associated with this illegal behavior. The existence of this external effect finds its origin in the fact that the road is used as a common. Several tools are available for policy makers to internalize those costly consequences and enforce traffic light obedience.

Although there is a sizable scientific literature dealing with RLR, researchers generally focus on the predictors of such behavior and the impact of the countermeasures from an engineering and psychological perspective. The economic approach is quite inexistent, so that it is impossible to determine the appropriate intervention and be insured the public policy is efficient. It neglects also the economic dimension of driver's choice. This entry proposes an economic approach which provides a new perspective for this issue and gives an account of the specialized literature.

An economic approach is required because red light violation as an external effect calls for a public intervention. However, the cost of intervention has to be proportional to the social cost of violations. In other words, the cost element constitutes the crucial dimension not only for framing countermeasures but also for understanding the offender's choice.

The first section provides an economic explanation of the need for traffic light at intersection through a game theory approach emphasizing upon the needs of cooperation and fairness. The second section deals with an analysis of RLR from the driver's choice perspective. The consequences of RLR are identified in section "Consequences of Traffic Light Violations". Section "Regulation of Red Light Violations" reviews the available possibilities for enforcing this safety rule and regulating efficiently the RLRs through the economist's lens.

Signalized Intersections

Cooperation and Safety at Intersections

Without traffic regulation, the situation of two drivers reaching the intersection from opposite directions can be represented as a noncooperative game (Table 1). If both players stop, they suffer a time loss ($-t$); if they both go, they suffer a large loss related to a road crash ($-a$), with $a > t > 0$. If they make different choices, the stopping driver suffers the time loss, while the other driver breaks even. In this game, the Pareto optimum is reached when drivers take different strategies.

In that case, there is no dominant strategy. If the adverse driver is expected to stop (resp. go), the best decision is to go (resp. stop). The absence of unique equilibrium can lead to a suboptimal situation (the crash or the mutual stop). By forcing one driver to stop, traffic light can be considered as an exogenous intervention that imposes one Pareto optimal equilibrium. Then traffic light can find a justification from the economic perspective by making cooperation possible between drivers. Here a better cooperation is related with traffic safety.

Traffic Management: Fairness and Sustainable Cooperation

Different types of traffic regulation are possible for solving the coordination problem at intersection such as putting a stop sign or giving priority to the right. However, the two aforementioned solutions systematically give priority to the same users. For the sake of fairness and efficiency of traffic regulation, traffic light alternates the two equilibria. Drivers from the non-priority roads are not systematically penalized, and waiting time is shared between drivers coming from opposite direction. Thus, traffic light ensures fairness and reciprocity among users. It is an important feature.

Traffic Lights Violations, Table 1 Game matrix of drivers at intersection

	Driver 2	
Driver 1	Go	Stop
Go	$-a, -a$	$0, -t$
Stop	$-t, 0$	$-t, -t$

Indeed, fairness and reciprocity can motivate cooperative behavior (Fehr and Schmidt 2006) and can therefore contribute to driver obedience to this traffic rule. It also makes it a sustainable one.

Traffic management is another main objective of traffic light regulation. Traffic lights are sometimes used for ramp metering, in order to make new entrants wait during congested periods. Others are used for regulating speed in urban areas. Such regulation aims at producing a smoother and calmer traffic among the road users and avoiding congestion. A driver pays a limited period of waiting time in order to enjoy larger gains with a reduced total driving time. Although such traffic management tools aim at reducing road crashes and time wastes, and ultimately the related social costs, not all drivers always abide to these rules. In a sense, RL violation can be considered as a free riding activity. The free rider would like to benefit from traffic safety and management without participating to its funding by obeying the rule. It is also a departure from the Pareto optimal equilibrium defined previously.

Traffic Light Violations

Despite of the related crash risk, red light running (RLR) is not an uncommon event. Retting et al. (1998) observed an average of three violations per hour on two urban intersections in Arlington. Carnis and Kemel (2012) reported very similar results from a field investigation for 24 traffic lights in Nantes (France) and showed that RLR characteristics vary with the type of sites, users, and contexts.

Individual Choice

A RLR is mainly an individual choice. Indeed, the classical school of economics of crime assumes that traffic offenders choose whether to violate the law or not by following utility maximization rules (Becker 1968). For red light violation, the elements of this type of decision can be presented in a decision matrix (Table 2). Stopping at the traffic light results in a sure time loss $(-Ct)$, whereas the consequence of a red light violation is uncertain and event contingent. It depends on the presence of another vehicle crossing the intersection and enforcement (a police officer or camera). If one of these events occurs while the driver decided to run the light, a cost is suffered: $(-Ca)$ for the crash cost and $(-Cf)$ for the cost of a fine, with $Ca > Cf > Ct > 0$. Ca can also include a penalty for causing the crash by red light violation.

Traffic Lights Violations, Table 2 Matrix of go/stop individual decision at red light

	No enforcement	Enforcement	
	No crossing vehicle	No crossing vehicle	Crossing vehicle
Stop	$-Ct$	$-Ct$	$-Ct$
Go	0	Cf	$-Ca$

There is no dominant strategy and the decision depends on users' beliefs and preferences. Decision under uncertainty is classically modeled by expected utility. This model assumes that decision makers assign subjective probabilities to events and subjective utilities to consequences and choose the alternative that maximizes the mathematical expectation of their utility. Normalizing the utility with $U(0) = 0$, a driver is expected to run the light if $U(-Ct) < pf \times U(-Cf) + pa \times U(-Ca)$, where pf (resp. pa) is the subjective probability of being fined (resp. responsible of a road crash). RLR is thus expected to vary across individuals and contexts depending on attitudes and perceived risks. This framework predicts also that RLR decreases when Pf, Pa, Cf, or Ca increases or when Ct decreases.

Empirical studies bring evidence for most of these predictions. The literature shows that increasing detection probability (by the mean of red light cameras, for instance) reduces RLR (Council et al. 2005). Moreover Carnis and Kemel (2012) show that violation rates are higher during night and low-traffic time periods when crash risk is lower.

The impact of red light duration on RLR was highlighted by Retting et al. (2008). When waiting time is too long, drivers fail to respect it. Guidelines recommend not having red light durations that exceed 2 min (CERTU 2010).

Carnis et al. (2012) report field data showing that most RLRs occur during the very first seconds of the red phase, when the violation is the most profitable in terms of avoided waiting time.

Heterogeneity is also expected between users, because of the diversity of individual preferences. Retting et al. (1999) compare authors of RLR accidents to those of other accident types. Males are overrepresented among this population. Red light violators are also younger and more likely to be intoxicated. Porter and England (2000) observed a relationship between RLR and safety-belt use. Propensity to abide to red light also depends on the vehicle type (Carnis and Kemel 2012).

Coping with Dilemma and Interactions Between Drivers

The decision to respect the rule must be taken in a very short amount of time. Indeed, drivers must make the go/stop decision within a few seconds. Because of the urgency dimensions of the decision and drivers' cognitive limits, illegal actions can sometimes be taken by mistake (Depken and Sonora 2009). Decision to go or stop at light also requires the driver to analyze the situation because the presence of closely following vehicles must be checked. If the decision to stop is likely to trigger a rear-end collision, decision to run the light must be taken. Consequently, illegal decision can be followed in particular situation for avoiding harm and costly consequences. Those aspects are not generally accounted for by the economics of crime framework that assumes that decision makers have time to choose, face clear-cut situations, and feature perfect cognitive capabilities.

The dilemma that faces the driver approaching light received an important attention in the literature (Elmitiny et al. 2010; Papaioannou 2007). The situation in which the driver is unable to stop safely or crossing the intersection at the green light is called the dilemma zone. The dilemma zone is related to the duration of amber light and the approaching drivers' speed. Shortening this time increases RLR because drivers are not averted that the light will switch. Increasing this time increases the dilemma zone and may increase the number of drivers running amber light. Drivers exceeding speed limits are more likely to run amber and red light. Therefore, the dilemma zone does not only puzzle drivers but network managers as well.

Decision to run or not the red light is not only individual, but it is also impacted by other drivers' behavior. The choice to commit a RLR takes into consideration the presence of other (preceding or following) drivers. For instance, drivers are more likely to run a light when a preceding driver did so (Elmitiny et al. 2010, p. 110). Observing that the preceding user runs the light may provide valuable information for decision that enforcement is low or nonexistent. Even though following behavior can be rational, it also creates risk of rear-end crash if the preceding driver decides to stop at the traffic light.

Consequences of Traffic Light Violations

Safety Consequences

Red light violation is a major concern for the policy makers because of the number of road crashes and victims involved (McGee and Eccles 2003). From the economist standpoint, road crashes are interpreted as an external effect related to the common use of the road network.

Moreover, the urban intersection implies mainly the involvement of vulnerable users (pedestrians, bicyclers, and motorcyclists). It means also the collision is characterized by a true asymmetrical dimension in terms of vulnerability between the involved users in a traffic collision (for instance, vehicle vs. pedestrian).

Large-scale studies evaluating the prevalence of RLR are not common. Retting et al. (1999) report that accidents occurring at intersections represent 27% of all injury crashes in the United States. Accidents due to RLR are however less frequent. Over the 1992–1996 period, RLR crashes represented 3% of all fatal crashes and 7% of injury crashes on urban roads. Compared to the prevalence of violations, these figures suggest that the collision probability in case of RLR is much lower than one could expect. According to the economic approach, violators may also decide to run red light when the traffic conditions and the

visibility minimize crash risk. Carnis et al. (2012) observed that 90% of violations occur in the first two seconds of the red phase, when all lights of the intersection are red.

Paradoxically, a sizable part of intersection crashes derive from red or amber light stopping. Rear-end accidents are indeed not infrequent at signalized intersection. Their number has been found to increase after traffic light camera deployment (Erke 2009). Another frequent type of accident occurring at signalized intersection relates to left turns. Wang and Abdel-Aty (2006) estimate that these accidents rank third after rear-end and angle crashes for 1531 intersections in the state of Florida.

Traffic Regulation Consequences

Traffic regulation is another major objective for installing traffic lights. Therefore, consequences in terms of generated congestion and the related time losses have to be assessed. Time losses can be generated by two types of traffic light violations. First, when the traffic light regulates road access, failure to respect the red light disturbs traffic flow and increases congestion. In this case, the contribution of the marginal violator to congestion is small, but the overall effect can be important when violations are numerous. Second, when RLR occurs during a dense traffic condition, the RL runner can be stuck in the middle of the intersection and can totally freeze traffic on all junctions.

A better respect of red light can help in limiting congestion and save environmental costs related to air pollution. We are not aware of any study evaluating the impact of RLR on time losses, nor environmental costs, due to increased congestion, even if they have to be taken into consideration from an economic perspective.

Regulation of Red Light Violations

Red light violation is a source of external effects. It generates a social cost (mainly associated with the crash costs (material damages and injuries)), which requires internalization. Internalization of this external effect calls for an intervention aiming at the reduction of costs borne by the victims. To mitigate the consequences related to those illegal behaviors, the policy maker defines and implements a public policy. This social regulation intervention can be achieved by two different categories of policies: enforcement and other interventions.

The Enforcement Policy

Enforcing the Highway Code

Becker's seminal works on the economics of crime show that illegal behavior can be mitigated by implementing an efficient policy of control and punishment (Becker 1968). Both the enforcer and the enforcement authority are concerned by the economic approach to crime. Efficiency of this enforcement policy requires taking into consideration the cost of intervention (respectively the relative costs of detection and punishment) and the social loss related to the harmful consequences of red light violations which could be reduced by deterrence. At the society level, it then becomes possible to determine an optimal deterrent policy associated with an optimal punishment (in terms of intensity of detection and severity of sanction) and an optimal number of violations. Consequently, it is rational from the economic perspective not to enforce all RLRs.

Different Techniques of Production

Different ways exist to enforce traffic light regulation. The traditional approach rests upon the manual detection of offenders by police officers, who monitor and intercept the offenders. This procedure is very costly in terms of time, because it requires a permanent supervision and numerous police officers to be able to catch the offender. In economic terms it is a labor-intensive technique of production. In practice, red light regulation was not especially enforced, because of its high unitary enforcement costs.

Since the mid-1980s, red light cameras (RLCs) have been replacing progressively the traditional enforcement method. This technique of detection can be considered as capital intensive and makes possible a systematic supervision of all the drivers, while minimizing the costs of labor intervention. Automation of traffic safety enforcement is a major trend of those last years, which has to be considered for understanding the spreading of such public programs.

When compared with the traditional approach, the RLC program appears as an efficient way for enforcing the regulation. It presents twofold economic advantages. It reduces substantially the cost of detection and punishment at a given level of traffic. The picture of the offender is automatically processed. The offender is identified through his license plate and receives his traffic infringement notice at home. It permits also to increase substantially the level of detection and punishment. Thousands or millions of tickets can be processed according to the limits of the computer system. In France, the number of RLR tickets was multiplied by 8 after the introduction of RLC. Introduction of RLC programs can be conceived as an innovation lowering the average cost of deterrence and making possible a stricter enforcement of red light regulation by generating scale economies. This cost killing effect explains probably why so many jurisdictions implemented such programs for securing the signalized road intersection (Carnis 2010).

Do RLCs Reduce Crashes?

While several contributions conclude to a positive contribution of RLC by reducing road injuries (Council et al. 2005; McGee and Eccles 2003), others show more debatable effects and question their impact. RLC would yield positive side effects with potential spillover impacts of RLC for other intersections and negative ones by increasing rear-end crashes and all category crashes (Hallmark et al. 2010; Vanlaar et al. 2014). However the gains associated with the reduction of right-angle injury crashes would largely compensate the costs related to the increase of rear-end crashes. More problematic are the recent conclusions of several contributions showing the insignificant impact of RLCs for reducing road crashes (Erke 2009; Høye 2013), contributions which were nevertheless criticized by other scholars putting in question their meta-analysis approach (Lund et al. 2009).

Cost-Benefit Evaluation is Needed

An economic approach to red light violation and regulation becomes particularly necessary when such a public intervention yields opposite and potential adverse side effects. It constitutes a prerequisite for concluding about the economic efficiency of such programs for reducing road injuries at signalized intersection. Proceeding to the economic assessment of RLC programs requires a comparison between advantages and costs. However, only few studies investigated the economic side of red light violations. More problematic is the finding of a careful literature review showing the quasi-generalized absence of economic assessment of RLC programs and rigorous evaluation of safety impacts, so that it is impossible to conclude that such programs are efficient and to determine the scope of the internalization policy (Langland-Orban et al. 2014). In fact, the present evaluative practices of RLC programs reflect both the complexity of evaluation process (non-replication of experiences in controlled laboratory conditions) and the costs of collecting and analyzing the data. It seems also to reflect that policy makers sometimes look for intervention whatever may be their impact or cost, when facing the risk of human injuries.

Public-Private Partnership for RLR Enforcement

The economic approach is particularly relevant when programs are not directly managed by governments. There are several procurement alternatives. Some of them could associate private operator, while some governments outsource the operation of the program (FHA 2005). The total or partial outsourcing of such social regulation activity raised some new issues concerning the possibility for contractor to manipulate the control activity and illegal use of the collected data (CSA 2002). Such situation is typically a principal-agent situation with asymmetrical information. Indeed, one agent is usually more informed than the other and can modulate its efforts. This contractual dimension emphasizes the necessity of a well-designed contract to be insured that private and public interests are aligned (Travis and Baxandall 2011). Indeed, while governments are more interested in maximizing their return in terms of safety impact (public safety hypothesis), the private firms are more concerned by the maximization of profit. Those considerations are quite important, because

it could influence the location of radars and their impact for public safety.

Another interesting issue is related to the different payment options for the contractor. Fixed-price payments, fixed monthly payments, per citations payments, payments depending on time worked and materials used, and mixed payments are alternative possibilities. However, it is not clearly determined which type of payment is the most advantageous for the government and the most efficient in terms of welfare for the society. Nevertheless there are strong incentive implications for the different agents, especially here for the policy maker and the contractor. Unfortunately this dimension was not investigated.

RLC programs have to be considered also from the institutional perspective (Carnis 2010). Outsourced programs, contractual dimensions, and financial considerations are important characteristics. A more recent trend fires on the RLC programs. More and more governments turned off their RLC because of uncertain impacts in terms of traffic safety already mentioned. Between 2011 and 2013, it is estimated that 200 RLC programs were turned off in the United States (Slone 2014). The policy maker is reluctant to let a program continue while it could increase rear-end injury accidents and potentially jeopardize human lives. Moreover the court system becomes less supportive of such control system: the judge dismisses charges more often because of erroneous readings and identification of some license plates by the program, which questions its reliability. Another consideration concerns the reduction of revenues associated with the RLC while the costs are increasing. Moreover some citizens assimilate RCL fines as a new tax imposed upon the road user. Garrett and Wagner (2009) concluded that sustained municipalities obey revenue motives concerning traffic enforcement and tickets. RLC would not be exclusively concerned with public safety.

Other Policies

Providing Drivers With Better Information

The drivers' decision of stopping or running the red light depends on beliefs and preferences. However most of the time the driver is not perfectly informed about the risks involved. Consequently, education and awareness campaign can play a useful role in providing accurate information related with the risk the driver faces when taking an adverse decision (FHA 2005). While education and awareness campaigns are conceived here as a provision for helping him in taking a correct decision, it presents a cost. Such campaigns have to be calibrated so that the costs and returns are in a same magnitude.

Nudging Drivers

Behavior change can also be achieved through nudging. This policy consists of framing the decision context. More precisely, the policy maker can design an environment that induces a promoted behavior. Thaler and Sunstein (2008) provide an illustrative example of such policy in the road safety field. The use of stripes can reinforce the visibility of a potential danger of a portion of a curve for instance (Thaler and Sunstein 2008, pp. 37–39). A review of possible applications of nudges to traffic safety policies is proposed by Avineri (2014). Regarding RLR, installing countdown systems can impact drivers' behavior and appears as a typical nudging intervention.

No Traffic Light, No Traffic Light Violation

Red light violation depends also on some Highway Code adaptations and road infrastructure context. For instance, in the United States, users are generally allowed to a right turn during red light as long as they leave priority to other vehicles. During low-traffic hours (e.g., night hours), traffic light can be turned out and the right of way applies, avoiding the unnecessary waiting time at the green light phase. In France, an experiment tests the impact of giving to the bicycle user the possibility to cross the intersection at a red light provided that the priority is given to the opposite coming vehicle.

Reframing the context of the driver decision can also require bringing some modifications to the road infrastructure. Engineer investigations showed the influence of the average daily traffic

volume, the number of traffic lanes, the left-turn lanes, and the speed limit regulation (Langland-Orban et al. 2014). A radical solution for regulating red light violation would consist of removing signalized intersection. In that case, RLR would disappear because the road infrastructure design makes them impossible. In some ways, it could constitute a kind of situational crime prevention approach. Concretely, it would consist of modifying the access to the road section through ramps or implementing roundabouts.

However such interventions can be very costly, especially in an urban context. Again a reasonable economic approach would consist of comparing costs and gains of different alternatives. This approach permits also to avoid the funnel approach by enlarging the problematic to other issues such as mobility and pollution considerations, emphasizing that red light violation prevention cannot be reduced only to public safety consideration.

Conclusion

The economic approach provides a consistent framework for understanding RLR. It is able to account for both users and policy makers decisions and highlight possible alternatives for intervention. It can also explain why it can be rational for a driver to commit a RLR under certain circumstances, but also why red light regulation is not enforced in some cases.

Economic variables are not only at play for explaining the way drivers choose in particular situations, but the economic consequences related to road crash and traffic congestion have also to be considered for understanding the role of traffic light. Economic valuation appears as a true alternative to the engineering perspective for understanding this issue and promotes different analysis and solutions.

Regulation of RLR is achievable and requires a calibrated enforcement policy. RLC program is a possible solution for enforcing traffic safety rules, but an economic approach is needed for designing correctly the public intervention, which could be assimilated to a particular productive process. Communication campaign, awareness program, and infrastructure modification are other available solutions. Nudging policy appears also as an interesting perspective that can be built upon a behavioral law and economic approach, providing new insights for designing traffic safety rules and enforcement policies.

References

Avineri E (2014) Nudging safer road behaviours, technical report. Afeka Center for Infrastructure, Transportation and Logistics

Becker G (1968) Crime and punishment: an economic approach. J Polit Econ 78:168–217

Carnis L (2010) A neo-institutional economic approach to automated speed enforcement systems. Eur Trans Res Rev 2(1):1–12

Carnis L, Kemel E (2012) Assessing the role of context in traffic light violations. Econ Bull 32(4): 3386–3393

Carnis L, Dik R, Kemel E (2012) Should I stay or should I go? Uncovering the factors of red light runnings in a field study. In: Proceedings of the 40th European transport conference, Glasgow, UK

CERTU (2010) Guide de conception des carrefours à feux, technical report. Centre d'Etudes sur les Réseaux, les Transports, l'Urbanisme et les constructions publiques

Council FM, Persaud B, Eccles KA, Lyon C, Griffith MS (2005) Safety evaluation of red-light cameras, US Department of Transportation, Federal Highway Administration, FHWA-HRT-05-048, pp 95

CSA (2002) Red light camera programs: Although they have contributed to a reduction in accidents, operational weaknesses exist at the local level, technical report. Bureau of State Audit

Depken C, Sonora R (2009) Inadvertent red light violations: an economic analysis, Mimeo, p. 32. belkcollegeofbusiness.uncc.edu.cdepken/P/redlights.pdf

Elmitiny N, Yan X, Radwan E, Russo C, Nashar D (2010) Classification analysis of driver's stop/go decision and red-light running violation. Accid Analy Prev 42(1): 101–111

Erke A (2009) Red light for red-light cameras? A meta-analysis of the effects of red-light cameras on crashes. Accid Analy Prev 41(5):897–905

Fehr E, Schmidt K (2006) The economics of fairness, reciprocity and altruism–experimental evidence and new theories'. In: Handbook of the economics of giving, altruism and reciprocity, Vol. 1, Elsevier, Amsterdam, pp 615–691

FHA (2005) Red light camera systems, operational guidelines, technical report. Federal Highway Administration
Garrett TA, Wagner GA (2009) Red ink in the rearview mirror: local fiscal conditions and the issuance of traffic tickets. J Law Econ 52(1):71–90
Hallmark S, Orellana M, McDonald T, Fitzsimmons E, Matulac D (2010) Red light running in Iowa. Trans Res Rec: J Trans Res Board 2182(1):48–54
Høye A (2013) Still red light for red light cameras? An update. Accid Analy Prevent 55:77–89
Langland-Orban B, Pracht EE, Large JT, Zhang N, Tepas JT (2014) Explaining differences in crash and injury crash outcomes in red light camera studies. Eval Health Prof 1–19 (Epub ahead of print)
Lund AK, Kyrychenko SY, Retting RA (2009) Caution: a comment on Alena Erke's red light for red-light cameras? A meta-analysis of the effects of red-light cameras on crashes. Accid Analy Prevent 41(4): 895–896
McGee H, Eccles K (2003) Impact of red light camera enforcement on crash experience, a synthesis of highway practice, technical report. NCHRP synthesis
Papaioannou P (2007) Driver behaviour, dilemma zone and safety effects at urban signalized intersections in Greece. Accid Analy Prevent 39(1):147–158
Porter BE, England KJ (2000) Predicting red-light running behavior: a traffic safety study in three urban settings. J Safety Res 31(1):1–8
Retting RA, Williams AF, Greene MA (1998) Red-light running and sensible countermeasures: summary of research findings. Trans Res Record: J Trans Res Board 1640(1):23–26
Retting RA, Ulmer RG, Williams AF (1999) Prevalence and characteristics of red light running crashes in the united states. Accid Analy Prevent 31(6): 687–694
Retting RA, Ferguson SA, Farmer CM (2008) Reducing red light running through longer yellow signal timing and red light camera enforcement: results of a field investigation. Accid Analy Prevent 40(1):327–333
Slone S (2014) Speed and red light cameras law, technical report. Capitol Research, The Council of State Governments
Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Haven
Travis M, Baxandall P (2011) Caution: red light camera ahead the risks of privatizing traffic law enforcement and how to protect the public, technical report. US PIRG Education Fund
Vanlaar W, Robertson R, Marcoux K (2014) An evaluation of Winnipeg's photo enforcement safety program: results of time series analyses and an intersection camera experiment'. Accid Analy & Prevention 62: 238–247
Wang X, Abdel-Aty M (2006) Temporal and spatial analyses of rear-end crashes at signalized intersections. Accid Analy Prevent 38(6):1137–1150

Transaction Costs

Wim Marneffe, Samantha Bielen and Lode Vereeck
Faculty of Applied Economics, Hasselt University, Diepenbeek, Belgium

Abstract

Transaction costs is a generic term referring to the costs of transacting through the market (e.g., search, information, contract, monitoring costs). They are applied with different meanings to organizational structures (e.g., vertical integration), market failures (e.g., externalities), institutional choices (e.g., promotion of clubs), and public choice (e.g., administrative burden).

Introduction

If markets operate as smoothly and efficiently as free market proponents and standard economic textbooks suggest, why, for instance, do firms hire workers on a permanent basis or are professions regulated? The answer is, to a large extent, transaction costs. In principle, every single activity (e.g., typing a letter, making a phone call) can be outsourced to the market. However, searching, finding, screening, contracting, and monitoring a secretary is extremely costly. These "transaction" costs of using the (labor) market can outweigh the proclaimed benefits of allocating productive activities (i.e., a letter, a phone call) through the market. Therefore, legal instruments are created (e.g., labor contract of unlimited duration) in order to avoid transaction costs.

Similarly, the regulation of professions (e.g., licensing of dentists) not only guarantees a standard quality, but also creates barriers of market entry. However, the oligopolistic costs that stem from these barriers are typically offset by the search, information, screening, contracting, and monitoring costs of using the free market of professional services, i.e., the transaction costs.

Transactions costs should always be considered in their historical context. The use, for instance, of labor contracts of unlimited duration is a way to reduce transaction costs outweighing monopolistic costs (i.e., barring other secretaries to type a letter or make a call). However, the emergence of temporary workers' agencies that offer flexible services is a market response to high transaction costs. What is an economic reason to avoid the use of the markets today may not be valid tomorrow.

The concept of transaction costs is probably among the most discussed topics in economics and has led to the emergence of an entire body of literature with numerous theoretical and empirical articles and books. Economists' understanding of transaction costs has continuously evolved since it was first introduced by Ronald Coase trying to explain the emergence of firms (Coase 1937). Currently, transaction costs are an essential part of economic analysis and frequently invoked to explain a plethora of institutional choices and behavior (Rao 2003). Unfortunately, the transaction cost literature is plagued by the use of different definitions. Therefore, this essay presents a consistent and coherent taxonomy of transaction costs.

Transaction Costs: The Prelude

Although Ronald Coase did not explicitly use the words "transaction costs" in his 1937 article on "The Nature of the Firm," he was the first to introduce the concept in order to explain the existence of firms instead of organizing economic activity through exchange of transactions across the market (Coase 1937). According to Coase, "we had a factor of production, management, whose function was to coordinate. Why was it needed if the pricing system provided all the coordination necessary?" (Coase 1993). Before Coase, creating a firm or using the market were viewed as alternative modes for coordinating production. At the time, the mainstream microeconomic assumption was that the use of the firm or the market for production was a given, not a choice (Williamson 2010).

Coase was the first to observe that firms arise because there are substantial costs involved in using the price mechanism of the market. At first, he was not very explicit about the meaning of these transaction costs. He did not provide a clear definition, but described transaction costs as the cost of "discovering what the relevant prices are" or "negotiating and contracting costs" (Coase 1937). Back in 1937, Coase did not fully grasp his own accomplishment and was merely trying to reveal the weaknesses of the dominant Pigouvian analysis of the divergence between private and social products (Coase 1993). In the late 1950s, research on market failures started, notably after the article of Samuelson on "The Pure Theory of Public Expenditures" (Samuelson 1954). At the time, few authors (e.g., Coase, Buchanan, Calabresi) considered externalities not as a market failure, but pointed out that under certain conditions "harmful effects" are not problematic for the efficiency of market mechanisms (Marciano 2011).

Transaction Costs: The Sequel

In 1960, Coase published another article on the "Problem of Social Cost" which is unmistakably one of the most cited articles in the economic and legal literature and helped launch the economic analysis of the law (Medena 2011).

In analyzing externalities (or "social cost issues" as they were called back then), Coase used the famous example of a rancher whose cattle destroys crops on the land of a neighboring farmer. First, it is assumed that the cattle rancher is fully liable for harm caused "and the pricing system works smoothly," i.e., a zero transaction cost model. Under liability, the efficient level of both cattle and crops will be produced, either through damage payments from the rancher to the farmer or through compensatory payments from the rancher to the farmer to take land out of cultivation (if that is the least-cost solution). This way, external costs are fully internalized under the liability rule. The actual level of damage payments is determined by "the shrewdness of the farmer and the cattle-raiser as bargainers" (Coase 1960).

Second, Coase discusses a situation in which the rancher cannot be held liable for the damage caused by his cattle. In this case, the resulting level of production of both cattle and crops is also efficient, since it is in the farmer's own interest to pay the rancher to cut the size of his herd up till the point where the payment is less than the benefit from reduced crop damage. Similarly, the rancher is willing to cut the size of his herd if the farmer's payment at least equals the foregone benefit resulting from the herd reduction. Again, external costs are fully internalized and both cattle and crop outputs are equal at the levels under liability.

Coase concludes that the initial assignment of legal property rights has no impact on the efficient use of resources. As discussed in the previous paragraph, both situations liability and non-liability lead to an efficient level of production of both cattle and crops. However, this conclusion only holds when the pricing system works at zero transaction costs. This insight has become known as "the Coase theorem." Interestingly, it was Stigler who coined this term in 1966 (Stigler 1966). In case of multiple farmers, substantial transaction costs may arise. Under liability, each individual farmer sues the rancher, hence transaction costs are low and the efficient level of production is attained. In case of non-liability, however, each individual farmer faces three inefficient options: (1) do nothing and bear the costs of crop damage, (2) build a fence around the individual property which on aggregate may outweigh the costs of herd reduction, or (3) try to negotiate a herd reduction which may cause substantial transaction costs, such as gathering information, contacting and discussing with other farmers, and negotiating with the rancher. Putting aside the potential free-rider problem, the latter "transaction" costs, however, can be so substantial that the farmers resort to option one or two.

Later on, Coase (1960) clarified that "to carry out a market transaction, it is necessary to discover who it is that one wishes to deal with, to inform people that one wishes to deal and on what terms, to conduct negotiations leading op to a bargain, to draw up the contract, to undertake the inspection needed to make sure that the terms of the contract are being observed and so on." The costs that accompany these activities may hamper transactions that would have taken place if using the pricing system did not evoke such costs (Coase 1960). Furthermore, he suggests that the existence of externalities can partly be explained by the presence of transaction costs that are sufficiently large to prevent market-functioning mechanisms to internalize external costs. At the time, Coase criticized the neoclassical Pigouvian model for ignoring the existence of transaction costs. In 1993, Coase pointed out that his article had demonstrated "the emptiness of the Pigouvian analytical system" and helped to frame the discussion on externalities in a more realistic way (Coase 1993). The Coase theorem was initially met with a lot of skepticism by other scholars who condemned the underlying assumption of efficient markets as unrealistic and even "Utopian" (Blum and Kalven 1967).

Transaction Costs: Definitions

The publication of "The Problem of Social Cost" in 1960 did not lead to the immediate absorption of the idea of transaction cost reasoning in economic literature. In 1969 Kenneth Arrow defined transaction costs as "the costs of running the economic system" (Arrow 1969). But, it was not until 1985 that transaction cost reasoning became more widely known due to Oliver Williamson, who defined it as "the economic equivalent of friction in physical systems" (Williamson 1985). In an effort to put the concept of transaction costs into practice, Williamson explains: "Transaction cost economics is an effort to better understand complex economic organization by selectively joining law, economics, and organization theory. As against neoclassical economics, which is predominantly concerned with price and output, relies extensively on marginal analysis, and describes the firm as a production function (which is a technological construction), transaction cost economics (TCE) is concerned with the allocation of economic activity across alternative modes of organization (markets, firms, bureaus, etc.), employs discrete structural analysis, and describes the firm as a governance structure

(which is an organizational construction). Real differences notwithstanding, orthodoxy and TCE are in many ways complements – one being more wellsuited to aggregation in the context of simple market exchange, the other being more well-suited to the microanalytics of complex contracting and nonmarket organization" (Williamson 2008).

In turn, Barzel (1997) described transaction costs as "the transfer, capture, and protection of exclusive property rights." This is a rather narrow definition of transaction costs, yet often used by scholars from the property rights movement.

More recently, the scope of transaction costs was broadened, leading Challen (2000) to state that transaction costs include all costs associated with any allocation decision, including the costs of uncertainty. Stavins (1995) claimed that transaction costs are "ubiquitous" in market economies, since parties must find one another to transfer, communicate, and exchange information. Douglass North (1990) went even further and considered transaction costs as a part of production costs.

Nowadays, transaction costs are an essential part of mainstream economics and are being applied with different meanings to organizational structures (e.g., vertical integration), market failures (e.g., externalities), institutional choices (e.g., promotion of clubs), and public choice (e.g., administrative burden). Transaction cost is now a generic term referring to costs occurring when making a transaction in the market. Accordingly, transaction costs can be interpreted as "the costs of any activity undertaken to use the price system" (Demsetz 1997).

Towards a Classification of Transaction Costs

The discussion above concerning the definition of transaction costs clearly shows the need for a coherent and complete classification of transaction costs. Dahlman (1979) was one of the first scholars to propose a categorization of transaction costs. In accordance with Crocker (1971), he distinguished three types of costs: (1) search and information costs, (2) bargaining and decision-making costs, and (3) monitoring and enforcement costs. However, Milgrom and Roberts (1992) used another classification of transaction costs: on the one hand, costs stemming from information asymmetries and incompleteness of contracts among parties and, on the other hand, costs following imperfect commitments or opportunistic behavior of parties. Furthermore, Foster and Hahn (1993) brought some new elements into the discussion and emphasized the distinction between direct financial costs (of engaging in trade), costs of regulatory delay, and indirect costs (associated with the uncertainty of completing a trade). A more basic classification is made by Dudek and Wiener (1996) who included search, negotiation, approval, monitoring, enforcement, and insurance costs.

One of the most interesting classifications of transaction costs is the one by Furubotn and Richter (1997). They describe transaction costs as the costs of establishing, maintaining, adapting, regulating, monitoring, and enforcing rules as well as executing transactions. Their definition of transaction costs is "the costs of resources utilized for the creation, maintenance, use, change, and so on of institutions and organizations. [. . .] When considered in relation to existing property rights and contract rights, transaction costs consist of the costs of defining and measuring resources or claims, plus the costs of utilizing and enforcing the rights specified. Applied to the transfer of existing property rights and the establishment or transfer of contract rights between individuals (or legal entities), transaction costs include the costs of information, negotiation, and enforcement." The authors identify three sorts of transaction costs: the costs of using the market (market transaction costs), the costs of exercising the right to give orders within the organization (managerial transaction costs), and the costs of running and adjusting a political system (political transaction costs). Each of these three categories comprises both fixed transaction costs (setup costs for institutional arrangements) and variable transaction costs (dependent on the number of transactions). Furthermore, it should be noted that the authors make a useful distinction between ex ante (e.g.,

search and information costs) and ex post (e.g., monitoring and enforcement costs) transaction costs. Thus, the Furubotn and Richter classification integrates the costs of using the market as mentioned by Coase, the managerial costs identified by Williamson, and the institutional costs put forward by North.

Unfortunately, the Furubotn-Richter taxonomy has one serious drawback. In their effort to provide a comprehensive classification, they associate "transaction costs" with the use of markets as well as regulations, which undermines the very meaning of the concept. Furubotn and Richter (1997) rightly claim that regulation entails costs. However, these costs are precisely the opposite from the (Coasean) transaction costs, which refer to the use of the market and thus terminologically unsound. Moreover, the policy goal of regulations is precisely to reduce transaction costs. Adding to the confusion, the OECD in 2001 also made the distinction between non-policy-related transaction costs (in which parties incur costs of voluntary market transactions) and policy-related transaction costs (resulting from the implementation of public policy).

It is clear that the distinction between transaction costs and regulatory costs should be strictly observed. Therefore, Marneffe and Vereeck (2011) suggest that the term "regulatory costs" is exclusively used to refer to the costs of interfering in, correcting, or barring the use of markets. They recommend to short-term the policy-related costs of regulation as "regulatory costs" and non-policy-related costs of using markets as "transaction costs."

Cross-References

- ▶ Coase, Ronald
- ▶ Coase Theorem

References

Arrow K (1969) The organization of economic activity: issues pertinent to the choice of market versus non-market allocation. In: US Joint Economic Committee (ed) The analysis and evaluation of public expenditure: the PPB system, vol 1. Government Printing Office, Washington, DC, pp 59–73

Barzel Y (1997) Economic analysis of property rights. Cambridge University Press, Cambridge

Blum WJ, Kalven H (1967) The empty cabinet of Dr. Calabresi auto accidents and general deterrence. Univ Chicago Law Rev 34(2):239–273

Challen R (2000) Institutions, transaction costs and environmental policy: institutional reform for water resources. Edward Elgar, Cambridge, MA

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

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

Coase R (1993) Law and economics at Chicago. J Law Econ 36:239–254

Crocker TD (1971) Externalities, property rights and transaction costs: an empirical study. J Law Econ 14:445–463

Dahlman CJ (1979) The problem of externality. J Law Econ 22:141–162

Demsetz H (1997) The firm in economic theory: a quiet revolution. Am Econ Rev 87:426–429

Dudek DJ, Wiener JB (1996) Joint implementation and transaction costs under the climate change convention. Restricted Discussion Document ENV/EPOC/GEEI (96)1. OECD, Paris

Foster V, Hahn RW (1993) Emission trading in LA: looking back to the future. American Enterprise Institute, Washington, DC

Furutbotn EG, Richter R (1997) Institutions and economic theory: the contribution of the new institutional economics. University of Michigan Press, Ann Arbor

Marciano A (2011) Ronald Coase, "The problem of social cost" and The Coase theorem: an anniversary celebration. Eur J Law Econ 31:1–9

Medena SG (2011) A case of mistaken identity: George Stigler, the problem of social cost and the Coase theorem. Eur J Law Econ 31:11–38

Milgrom P, Roberts J (1992) Economics, organization, and management. Prentice-Hall, New York

North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge

OECD (2001) Domestic transferable permits for environmental management: design and implementation. OECD Proceedings, Paris

Rao PK (2003) The economics of transaction costs: theory, methods, and applications. Palgrave MacMillan, New York

Samuelson P (1954) The Pure Theory of Public Expenditure. Review of Economics and Statistics 36:387–389

Marneffe W, Vereeck L (2011) The meaning of regulatory costs. European Journal of Law and Economics 32:341–356

Stavins R (1995) Transaction costs and tradable permits. J Environ Econ Manag 29:133–148

Stigler G (1966) The theory of price, 3rd edn. Macmillan, New York

Williamson O (1985) The economic institutions of capitalism. Free Press, New York

Williamson O (2008) Handbook of new institutional economics part I. Springer, Berlin/Heidelberg
Williamson O (2010) Transaction cost economics: the natural progression. Am Econ Rev 100:673–690

Transferable Discharge Permits

Cornelia Ohl
HA Hessen Agentur GmbH, Transferstelle für Emissionshandel und Klimaschutz, Wiesbaden, Germany

Abstract

There are different approaches for dealing with water, air, and soil pollution, for example, the prescription of an emission or immission standard by environmental law or pollution reduction by economic instruments like *transferable discharge permits* (TDP). The idea of TDP is to control the pollution level in the environmental media by economic incentive setting (see, e.g., Tietenberg T, Lewis L (2014) Environmental & natural resource economics, global edition, 10th edn. Pearson Higher Education; for an introduction in economic theory). This requires the setting of a critical threshold level for pollution control – e.g., a safe minimum standard – and the breakup of this standard ("cap") in permits that can be traded on a market.

The most prominent application of TDP in Europe is the trading of CO_2 emissions in selected sectors of the economy (EU-ETS; see, e.g., Ellerman D, Convery F, de Perthuis C (2010) Pricing carbon: the European Union emissions trading scheme. Cambridge University Press, Cambridge, UK/New York (also published in French: Le Prix du Carbone: Les enseignements du marché du carbone, London: Pearson, 2010); Endres and Ohl (Eur J Law Econ 19:17–39, 2005)). It can be seen as a flagship approach for a European-Union-wide harmonization of environmental laws and regulations for emission control by economic incentive setting.

Synonyms

Cap-and-trade approach; Emissions trading; Tradeable discharge permits; Trading of allowances

What Is It About? Proposal for an Alternative Heading: The Idea of Transferable Discharge Permits

There are different approaches for dealing with water, air, and soil pollution, for example, the prescription of an emission or immission standard by environmental law or pollution reduction by economic instruments like *transferable discharge permits* (TDP). The idea of TDP is to control the pollution level in the environmental media by economic incentive setting (see, e.g., Tietenberg and Lewis 2014 for an introduction in economic theory). This requires the setting of a critical threshold level for pollution control – e.g., a safe minimum standard – and the breakup of this standard ("cap") in permits that can be traded on a market.

The most prominent application of TDP in Europe is the trading of CO2 emissions in selected sectors of the economy (EU-ETS; see, e.g., Ellerman et al. 2010; Endres and Ohl 2005). It can be seen as a flagship approach for a European-Union-wide harmonization of environmental laws and regulations for emission control by economic incentive setting.

Market Design by Environmental Legislation

The standard can be fixed on different backgrounds, on social welfare considerations, aspects of human health, and nature protection, among others. In reality, we often find a mixture of different factors including political, economic, and natural science considerations. In any case, the standard is to split into permits which with regard to the regulated subject and area allow a certain amount of pollution and are valid for a certain time period. This poses questions of monitoring and measurement.

Pollution is usually a by-product of valued goods and services which make the measurement tricky, especially in cases where direct measurement is impossible and indirect calculations or assessments are required. To ensure that the assigned amount can be traded, the certified metric needs to mean the same for any regulated body irrespective of the type of good or service it provides. For this reason, clear regulations on monitoring and measurement are necessary. It is also to ensure that the number of permits certifies the critical threshold value set by the regulating authority.

In a further step, the regulating body is to issue the permits either free of charge or by selling them for a fixed price or through bidding or auctioning mechanism. Moreover, to guarantee that the polluters perform with the standard, reporting on the polluting activities is essential. It calls for the establishment of an accounting system that can be verified by independent experts. This supports proving performance with monitoring and measurement rules and individual compliance, i.e., if the polluters keep their polluting activity within the limit allowed by the number of permits they hold. If the polluting behavior is inconsistent with the assigned amounts, an enforcement mechanism is needed that makes the polluter stick to its obligations, for example, a penalty fee in combination with a belated reduction of excessively released pollutant.

Incentives for Permit Trading

Keeping within the limits of a standard frequently requires measures for emission/immission reductions. These measures raise different costs for the regulated bodies. To keep them at minimum, economists argue for the establishment of a market where the permits can be traded.

Although the market does not guarantee that the polluters with the lowest cost actually reduce the pollutant – the polluter is generally free to decide whether to buy permits or change the polluting behavior – it nevertheless encourages minimizing the cost of pollution control. If polluters maximize profits, they will take advantage of the cheapest way of pollution reduction. The polluter thus has incentive to compare the market price of a permit with the individual reduction costs. As long as the cost for buying permits is lower than the cost of measures for pollution control, the polluter shows a demand for permits. On the other hand, polluters with low control costs are willing to reduce their pollution level in order to sell permits on the market, at least as long as the price they receive for the permit is higher than the individual reduction costs. This mechanism ensures that the standard is enforced at minimal costs.

Despite this advantage, the market approach is often criticized. One argument is that polluters with high reduction costs lose incentives for pollution control and let others do the job. The question, however, is whether society is concerned of the pollutant to stay within the prescribed limit or whether the concern is on who is responsible for taking measures. If the goal is pollution control, the market-based approach has clear advantages in terms of both social welfare and environmental protection: it first of all draws the focus on pollution reduction to the desired extent and second to incentive setting for cost minimization. The question of responsibility, nevertheless, can be addressed by selecting the group of polluters having to perform with the environmental regulation.

Applications and refinements of TDP as well as further criticism are found in the literature (e.g., see Hansjürgens et al. 2011; OECD 2004).

Cross-References

- ▶ Anticommons, Tragedy of the
- ▶ Coase and Property Rights
- ▶ Economic Efficiency
- ▶ Emissions Trading
- ▶ Law and Economics
- ▶ Market Definition

References

Ellerman D, Convery F, de Perthuis C (2010) Pricing carbon: the European Union emissions trading scheme. Cambridge University Press, Cambridge, UK/New York (also published in French: Le Prix du Carbone: Les enseignements du marché du carbone, London: Pearson, 2010)

Endres A, Ohl C (2005) Kyoto, Europe? – an economic evaluation of the European Emission Trading Directive. Eur J Law Econ 19:17–39
Hansjürgens B, Antes R, Strunz M (2011) Permit trading in different applications. Routledge, New York
OECD (2004) Tradeable permits: policy evaluation, design and reform. OECD Publishing, Paris. https://doi.org/10.1787/9789264015036-en
Tietenberg T, Lewis L (2014) Environmental & natural resource economics, global edition, 10th edn. Pearson Higher Education, Universal Free E-Book store

Further Reading (documents on EU-ETS)

Commission Decision 2006/780/EC of 16 November 2006 on avoiding double counting of greenhouse gas emission reductions under the Community emissions trading scheme for project activities under the Kyoto Protocol pursuant to Directive 2003/87/EC of the European Parliament and of the Council [Official Journal L 316 of 16 November 2006]
Commission Decision 2007/589/EC of 18 July 2007 establishing guidelines for the monitoring and reporting of greenhouse gas emissions pursuant to Directive 2003/87/EC of the European Parliament and of the Council [Official Journal L 229 of 31.8.2007]
Commission Regulation (EU) No 1031/2010 of 12 November 2010 on the timing, administration and other aspects of auctioning of greenhouse gas emission allowances pursuant to Directive 2003/87/EC of the European Parliament and of the Council establishing a scheme for greenhouse gas emission allowances trading within the Community [Official Journal L 302 of 18.11.2010]
Directive 2003/87/EC of the European Parliament and of the Council of 13 October 2003 establishing a scheme for greenhouse gas emission allowance trading within the Community and amending Council Directive 96/61/EC

Transition Economies: Rule of Law and Credible Commitment

Rosolino A. Candela[1] and Ennio Piano[2]
[1]Political Theory Project, Department of Political Science, Brown University, Providence, RI, USA
[2]Department of Economics, George Mason University, Fairfax, VA, USA

Abstract

Must the rule of law spring upon us solely by accident and force, or can it also emerge as a product of political reflection and choice? As the lessons from the political and economic transition of centrally planned economies in Europe and Asia illustrate, the transitional movement requires an element of political reflection and deliberate choice, specifically the establishment a binding and credible commitment to limits on state action. In order to do so, credible commitment to the rule of law must be signaled *exante* by political reformers if reform efforts are to be successful, from which economic development follows *expost*. The historical record shows, however, that efforts to establish the rule of law in transition economies have been mixed. This chapter explains why this is the case, specifically by comparing Russia and China as case studies of transitional political economy.

Introduction

Must the rule of law spring upon us solely by accident and force, or can it also emerge as a product of political reflection and choice? In a previous chapter of this encyclopedia, we discussed the role in which interjurisdictional competition between states and intrajurisdictional competition between interest groups constrained governments to obey its political commitments to enforce property rights, in which the rule of law emerged as a by-product of this process. However, as the lessons in the political and economic transition of centrally planned economies show, the transitional movement requires an element of political reflection and deliberate choice (see Hayek 1973: 45; Wagner and Runst 2011), specifically the establishment of a binding and credible commitment to limits on state action (Boettke 2009). In order to do so, credible commitment to the rule of law must be signaled *ex ante* by political reformers if reform efforts are to be successful, from which economic development follows *expost*. The historical record shows, however, that efforts to establish the rule of law in transition economies have been mixed. This chapter explains why this is the case, specifically by comparing Russia and China as case studies of transitional political economy.

The Problem of Credible Commitment

There lies a fundamental dilemma in every attempt to establish the rule of law in those economies undergoing a transition away from national economic planning: how do we resolve the paradox of governance? In order to create the conditions conducive to exchange and capital accumulation, governments must be strong to enforce property rights and contractual arrangements. Doing so enables individuals to devote less effort and resources preventing "first-level predation," meaning predation by private actors, and specialize in productive entrepreneurial activities. Eliminating first-level predation, however, introduces the problem of "second-level predation," or predation by public actors (Boettke 2009: 47; see Acemoglu and Johnson 2005). Governments that are strong enough to enforce private property rights and prevent private predation are also strong enough to break any constraints to such precommitments in the future. Public predation can manifest itself in several ways, including the outright confiscation of wealth, taxation, and inflationary finance through currency debasement (Montinola et al. 1995: 54). The cost of such predation in terms of economic development is "evasive entrepreneurship," which includes the expenditure of resources evading the legal system (Coyne and Leeson 2004: 57), via tax evasion, bribery, or withholding efforts to invest in physical and human capital. If government is given enough power to enforce the rules necessary to maintain an economy, then it must also credibly commit to honor such rules.

Why is it difficult for government officials to credibly commit to the rule of law? This can be explained by the absence of a "political Coase theorem" (Acemoglu 2003). The Coase theorem states that when private property rights are well defined and transactions are low, individuals will bargain to an efficient outcome, irrespective of the initial assignment of property rights (Coase 1960; Stigler 1966: 113). Implicitly, the Coase theorem assumes that parties to an exchange are full residual claimants and will therefore bear the full costs and benefits of the outcomes of their decision-making. Therefore, if one party to an exchange breaks an agreement, or if both parties are disputing an agreement, reputational mechanisms or the state is available to enforce contractual agreements. Failure to abide by an agreement will result in a loss of reputation in future trading periods and/or state punishment. The Coase theorem applied to the context of political decision-making, however, is quite different. The logic of political decision-making is to concentrate benefits on well-informed and well-organized interest groups that represent a politician's constituency and disperse costs on the masses of the ill-informed population (Boettke 1993: 7). Such a logic prevails because the governmental official who promises institutional reform to their citizenry will not bear the full costs of reneging on such a promise. Whereas in the previous example, a private actor faces costs in terms of state punishment and loss of reputation, government officials only bear a loss in reputation. However, even these costs are not fully concentrated on the particular public actor, but are spilled over onto future generations of public actors, who face an increasingly distrustful citizenry, making it more difficult in future "trading periods" to implement transitional reforms.

In order to overcome the expectations that the citizenry hold with regard to credibly committing to the rule of law, governments must signal their willingness to tie their hands *exante* against using discretion to unforeseen economic circumstances *expost*. Signaling refers to the transmission of information that is costly for a sender to emit, but is "cheap" for another party to receive, in this case government officials being the senders and the citizenry being the recipients. Without a credible commitment to the rule of law via signaling, "it is of course perfectly rational for private agents to discount announcements of future policy reforms – or assurances of the continuation of present reforms" (Rodrik 1989: 757). Rodrik argues that any successful institutional reform will have a bias towards "overshooting" in its signaling strategy. This implies that there will be inverse relationship between policy overshooting and credibility: the more severe the credibility gap, the greater the policy must overshoot in order to send the appropriate signal and overcome

the problem of credible commitment. For example, a central banker who wishes to credibly commit to sound monetary policy could adopt a rule that pegs the exchange rate, allowing citizens to convert the domestic currency into foreign currency at a preannounced rate. However, a central bank with a long record of inflationary policy will be sending a noisy signal, especially if devaluation is expected in the face of speculative attacks. In order to overcome the gap of credibly committing to currency debasement, policymakers may have to overshoot even further by adopting a more binding constraint that eliminates its discretion completely. Examples include dollarization or the private-issuance of competitive currencies (Selgin and White 2005; see also White 2010). Therefore, establishing a credible commitment to the rule of law not only requires sending a strong signal to execute institutional reform, but also establishing a binding constraint that eliminates political discretion. Once this takes place, economic and political transition can move into a wealth-creating path.

The Rule of Law and Transitional Political Economy

Fundamentally, transitional political economy entails an institutional change in the manner in which property rights are structured and governed, both in terms of formal government enforcement and informal institutional arrangements, such as customs, traditions, and norms. Broadly speaking, property rights refer to social relationships that guide expectations about the use of scarce resources (Furubotn and Pejovich 1972). Political and economic transitions are intertwined by changes in the de facto structure of property rights. The transition to a private property rights-based economy, based on freedom of exchange and freedom of contract, leads to a reallocation of wealth previously based on political connection and patronage. Respect for private property under the rule of law promotes political freedom because it separates economic power from political power and in this way enables the one to offset the other. It enables economic strength to be a check to political power rather than its reinforcement (Friedman 1962 [2002]: 15). Such an institutional transition will imply transaction costs that are not only economic and political, but cultural and historical as well.

Implementation of the rule of law in transitional economies implies the elimination of existing political privileges, specifically de facto control over the use of resources. Therefore, economic and political transition will generate a massive redistribution in income. This in turn will imply the opposition of a large class of government officials and military personnel whose rank and privileges will be threatened by transition to market economy under the rule of law. Theoretically, this problem could be eliminated by paying those individuals the present discounted value of the income derived from holding political power, particularly through de facto control over state-owned firms and its resources. However, when the transitional costs associated with compensating the current benefactors of the existing system are greater than the associated welfare gains dispersed among the population, this will create what Gordon Tullock refers to as a "transitional gains trap" (Tullock 1975) that impedes the implementation of the rule of law.

The process of institutional transition from a socialist economy to a market economy is not only economic and political in nature, but also cultural (Pejovich 2003). Culture refers to the context where goals of individuals and the means to be employed by individuals are shaped and given meaning (Storr 2013: 54). The way in which property rights are perceived and "what constitutes an appropriate disposition of property, are all (partly) determined by culture" (Storr 2013: 32). Therefore, culture will affect the ability for governments to credibly commit to uphold the rule of law. Steve Pejovich has argued that "*the cultural differences between Central and East European countries are a major determinant of the magnitude of their respective transaction costs*" (emphasis original, Pejovich 2003: 352). Countries that have a "command culture," which perceives the exchange of property rights to be a zero-sum game, will face higher costs of credibly committing to the rule of law than a "culture of

exchange" that regards the exchange of property rights to be mutually beneficial (see Buchanan 1997: 95–101). This is because if the "gains from trade are seen as a redistribution of income rather than as rewards to innovators for creating new wealth," then "[s]tate authorities are more likely to impose price controls on producers and/or merchants who earn large profits than to seek ways to create incentives for others to emulate such individuals in competitive markets" (Pejovich 2003: 351). The costs, however, of institutional transition present potential profit opportunities to be monetized by institutional entrepreneurs to change the rules of game, namely, through the establishment of private property rights (Leeson and Boettke 2009; Li et al. 2006). The role of institutional entrepreneurship will be discussed in the next section.

Moreover, not only does culture matter, but ideas also matter as well, namely, through the footprint of history. For example, one of the obstacles that China encountered during the early stages of its transition to a market economy, beginning after 1978, was the "Illusion of 1957" (Cheung 1982 [1986]: 26–27). According to this illusion, the Chinese regarded capitalism to be worse than communism by equating "capitalism" with the cronyism and corruption of the rule of the Kuomintang prior to 1949, which preceded a period of economic recovery under "communism" between 1949 and 1957. "Thus within living memory the Chinese people equate capitalism with the Kuomintang *débâcle* and communism with the 'good years' of 1949–1957" (emphasis original, Cheung 1982 [1986]: 28). Such historical perceptions, however false and misleading they may be, may only create further institutional inertia towards a complete and credible transition to a market economy.

Given all of these difficulties inherent not only to the problem of credible commitment via signaling, but also the transaction costs inherent to institutional transition, how have some countries been able to transition to a market economy under the rule of law and achieve economic development? Peter Boettke outlines the steps that must be followed to establish a path towards successful political and economic reform: "Reform in real time must (1) start from the existing status quo, (2) unearth the *de facto* organizing principles of that status quo, (3) design a set of reforms which address the incentive and informational problems associated with that *de facto* system, and (4) send a clear high-quality signal that the proposed reforms are credible and commit the governance structure to the new system and in doing so close the gap for the *de jure* and *de facto* organizing system in the new regime" (emphasis original, Boettke 1999: 378). In the next section, we will show how this transition has unfolded under comparative systems of federal governance, using the transitional political economy of Russia and China as comparative case studies.

Russia and China: A Comparative Transitional Analysis

One way to explain this institutional transition is to compare the transitional path that Russia and China have taken. In doing so, we make a distinction between "market-preserving federalism" (Weingast 1995; Qian and Weingast 1997) and "cartel-federalism" (Wagner and Yokoyama 2013; Wagner 2016). Under a federalist system of governance, governance is divided into two levels of authority: a national level of authority and subnational level of authority of smaller local government entities. The distinction between market-preserving federalism and cartel-federalism is based on how the national level of governance sets the conditions of competition between subnational governments.

Under market-preserving federalism, the national government credibly commits to the rule of law, namely, by allowing the entry and exit of private-sector firms to compete with state-owned firms. As a result, the wealth created through private-sector firms within competing local jurisdictions will grow, eroding the relative value of the rents derived from the political control of state-owned firms. Therefore, with the growing extension of the market, the allocation of entrepreneurial talent becomes redirected towards productive activities, rather than unproductive activities, such as rent-seeking

(Murphy et al. 1991; Tullock 1967). However, decentralization and competitive, or market-preserving, federalism are not synonymous. For example, the Soviet Union was "decentralized" in the sense that policy was administered by local governments, but local governments lacked political autonomy and local authority over the economy (Montinola et al. 1995: 57).

Under cartel-federalism, the national government creates a framework of collusion among subnational political entities. Like a cartel of firms, subnational governments collude to act as a collective monopoly, with the government acting as the de facto enforcer against attempted "chiseling" by local governments. In effect, cartel federalism prevents the erosion of rents derived from political control of state-owned firms and resources, namely, by obstructing the introduction of pro-growth policies among local governments that would encourage the growth of the private sector of the economy. The different institutional arrangements adopted by Russia and China since the late 1980s are reflected in their respective economic trajectories. Ten years after the transition reforms, Russia's GDP per capita had *fallen* by almost 30% (Leeson and Trumbull 2006). The same figures had more than doubled in China over the same period of time (World Bank 2017).

Whereas Russia's transition can be characterized by cartel federalism, China's transition followed more closely a model of market-preserving federalism. Under a model of market-preserving federalism, China's de jure changes were far less important that the de facto changes that emerged spontaneously among competing local jurisdictions. However, the sequence of events which we describe should not undercut "the essential element in political and economic transitions of post communism – the establishment of a binding and credible commitment to liberal limits on state action" (Boettke 2009: 43). While the emergence of private property in China was not designed by government policy, this bottom-up reform would not have flourished without a credible commitment by the state not to obstruct the de facto changes in property rights. After the rise of Deng Xiaoping in 1978, the "first priority under the new economic policy was agriculture" (Coase and Wang 2012: 157). However, agricultural reform was not initiated by the de jure privatization of farmland. Although land is still formally owned by the state, the introduction of a "household responsibility system" has led to a de facto reallocation of these rights to local stakeholders.

Under this system of landownership, peasants are allowed to lease an exclusive plot from the government, leaving the responsibility contract holder to grow and sell crops of their choice. By 1982, these responsibility contracts began to be traded among peasants, and such transfers became formally permitted by the government in 1983 (Cheung 1982 [1986]: 66). The household responsibility system arose in 1978, initially out of the institutional entrepreneurship of Yan Junchang, who was a villager in Xiao Gang, a poverty-stricken village in Anhui province of China. As Li, Feng, and Jiang describe the account, "[s]truggling to escape absolute poverty, on November 24, 1978, he and 17 other farmers signed a secret agreement to divide up the land and let each household work by itself, running the risk of jail sentences. They had the implicit support of local reform-minded officials. One year later, their innovation proved to be a big success: the total grain in production was equal to the sum of production over the previous five years" (2006: 245; see also Coase and Wang 2012: 47). This de facto privatization of property in agriculture occurred simultaneously with the rise of commerce in special economic zones (SEZs), which was first established in the Guangdong province on August 26, 1980, to attract foreign capital and investment (Coase and Wang 2012: 62).

It is no accident that the rise of the household responsibility system and SEZs coincided with institutional reforms that limited state action in terms of taxation and regulation within Chinese provinces. Starting in 1980, China instituted a fiscal revenue-sharing system between the provincial governments and the central government. According to this fiscal arrangement, revenue income in each province is divided between a fixed shares of revenue, which is remitted to the central government, allowing the remaining share of tax revenue to remain within the local

jurisdiction (Montinola et al. 1995: 63). Moreover, the Communist Party has retained authority to appoint and dismiss governments according to their ability to foster pro-growth policies. As an added incentive, and perhaps "as an ultimate prize, the governors whose regions perform well have been brought into the national government in Beijing" (Blanchard and Shleifer 2001: 175). This choice of institutional design addressed the incentive and informational problems associated with that de facto system, namely, by incentivizing and selecting for those local officials whose interests would be aligned with the pursuit of pursue pro-growth policies that encourage increased labor productivity and capital accumulation.

Although China remains undemocratic politically speaking, its ability to achieve rapid growth since 1978 could not have occurred without a credible commitment to limits on political discretion and public predation, without which property rights in land, labor, and capital would not have emerged to foster economic development. The transitional political economy of Russia, however, can be characterized by cartel federalism. Although Russia has held democratic elections and has engaged in a massive privatization scheme, such de jure reforms have not accompanied de facto reforms. This has occurred because of the failure of political officials to signal a credible commitment to the rule of law.

The Russian experience differed drastically from China's. Boettke (1993, 1995) argues that the failure of the attempts to reform the Russian economy (first under Gorbachev and then under Yelstin) is due to that the lack of an effective rule of law. Absent the rule of law, Russian citizens had to predict whether the government would go through with reform or, instead, reverse back to centralized control of the economy and social and political life. This prediction was rooted on the historical experience with economic reform in Russia since the time of Lenin: every time the Soviet government had announced a movement towards the devolution of economic and political decision-making, this decision was then reversed in the span of years and even months. With this reversion often came the persecution of those who had taken advantage of the new economic opportunities. It is not surprising that the Russian people did not accept the promises of the Gorbachev government at face value. Moreover, the behavior of the regime was itself sending mixed signals to the population, making the unraveling of the "reform game" even more likely. The three major pieces of legislation under *Perestroika* (the Law on Individual Enterprises, the Law on State Enterprises, and the Law on Cooperatives) aimed at increasing the competitiveness of the Soviet economy were all radically modified during the (brief) period of their implementation. "Despite the rhetoric and promise of these laws," Boettke writes, "they contained contradictions and ambiguities that prevented them from achieving the objectives of economic reform. Furthermore, they failed to convey any binding commitment on the part of the Gorbachev regime to true market reform. From 1985 to 1991, Gorbachev introduced at least ten major policy packages for economic reform under the banner of perestroika, *but not a single one was fully implemented*" (emphasis added, 1995 [2001]: 166).

As Russia began its process of privatization during the 1990s, it did so without a complete transition of its political institutions, within which privatization takes place. Without political reform of the institutions within which political decision-making takes place, the incentives that politicians faced during the transition period remained the same. As Milton Friedman points out, "Russia privatized but in a way that created private monopolies, private centralized economic controls that replaced government's centralized controls. It turns out that the rule of law is probably more basic than privatization. Privatization is meaningless if you don't have the rule of law. What does it mean to privatize if you do not have security of property, if you can't use your property as you want to?" (Friedman 2002: viii). The inconsistency between the spirit and the letter (and application) of the reform plan only added to the skepticism of the Russian people. If the past behavior of the Soviet government suggested that its stated goals were inauthentic, its present behavior only confirmed these suspicions.

Shliefer (1997) and Blanchard and Shleifer (2001) identify another cause of Russia's reform failure, namely, its fiscal institutions, which contrast with that of China. Given the lack of the rule of law, the introduction of a decentralized political system is unlikely to lead to the desired outcome. Unlike in China, where tax revenues are raised locally, this encourages local political officials to expand their tax base by fostering economic growth. The political context in Russia, however, was best characterized by cartel federalism, since "local government revenues comes from their share in taxes collected by the central government. Moreover, while this share is in theory fixed, in practice it is negotiated. Regional governments negotiate with Moscow, and local governments negotiate with regions" (Shliefer 1997: 403), creating a collusive rather a competitive environment between regions and disincentivizing the institutionalization of pro-growth policies. Indeed, while Russia's transition has been characterized by greater de jure reforms than China, there have been relatively less de facto reforms, precisely because of a failure to credibly commit to changing the institutional incentive structure within which economic and political decision-making takes place. As Coase remarked in his Nobel Prize Address on the heels of transition of Eastern and Central Europe, "[t]hese ex-communist countries are advised to move to a market economy, and their leaders wish to do so, but without the appropriate institutions no market economy of any significance is possible" (1992: 714).

Conclusion

The problem of economic transition is fundamentally institutional in nature. For economic and political reforms to be successful in transitional economies, this entails a credible commitment to the rule of law. Absent this commitment, aligning incentives between political and economic actors becomes extremely difficult, condemning the transition process to a likely failure. The experience of postcommunist Russia and postreform China provides evidence for this hypothesis.

Cross-References

- ▶ Capitalism
- ▶ De Jure/De Facto Institutions
- ▶ Development and Property Rights
- ▶ Fiscal Federalism
- ▶ Hayek, Friedrich August von
- ▶ Institutional Economics
- ▶ Political Economy
- ▶ Privatization
- ▶ Rule of Law and Economic Performance
- ▶ State-Owned Enterprises

References

Acemoglu D (2003) Why not a political coase theorem? Social conflict, commitment, and politics. J Comp Econ 31:620–652

Acemoglu D, Johnson S (2005) Unbundling institutions. J Polit Econ 113(5):949–995

Blanchard O, Shleifer A (2001) Federalism with and without political centralization: China versus Russia. IMF Staff Pap 48:171–179

Boettke PJ (1993) Why perestroika failed: the politics and economics of socialist transformation. Routledge, New York

Boettke PJ (1995/2001) Credibility, commitment and soviet economic reform. In: Boettke PJ (ed) Calculation and coordination: essays on socialism and transitional political economy. Routledge, New York, pp 154–175

Boettke PJ (1999) The Russian crisis: perils and prospects for post-soviet transition. Am J Econ Sociol 58(3):371–384

Boettke PJ (2009) Institutional transition and the problem of credible commitment. Annu Proc Wealth Well Being Nation 1:41–52

Buchanan JM (1997) Post-socialist political economy: selected essays. Edward Elgar, Lyme

Cheung SNS (1982/1986) Will China go 'Capitalist'?: an economic analysis of property rights and institutional change, 2nd edn. Institute of Economic Affairs, London

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

Coase R (1992) The institutional structure of production. Am Econ Rev 82(4):713–719

Coase R, Wang N (2012) How China became capitalist. Palgrave Macmillan, New York

Coyne CJ, Leeson PT (2004) The plight of underdeveloped countries. Cato J 24(3):235–249

Friedman M (1962/2002) Capitalism and freedom, 40th anniversary edition. University of Chicago Press, Chicago

Friedman M (2002) Preface: economic freedom behind the scenes. In: Gwartney J, Lawson R (eds) Economic

freedom of the world 2002 annual report. The Fraser Institute, Vancouver
Furubotn EG, Pejovich S (1972) Property rights and economic theory: a survey of recent literature. J Econ Lit 10(4):1137–1162
Hayek FA (1973) Law, legislation and liberty, Vol. 1: Rules and order. University of Chicago Press, Chicago
Leeson PT, Boettke PJ (2009) Two-tiered entrepreneurship and economic development. Int Rev Law Econ 29(3):252–259
Leeson PT, Trumbull WN (2006) Comparing apples: Normalcy, Russia, and the remaining post-socialist world. Post-Soviet Aff 22(3):225–248
Li DD, Feng J, Jiang H (2006) Institutional entrepreneurs. Am Econ Rev 96(2):358–362
Montinola G, Qian Y, Weingast BR (1995) Federalism, Chinese style: the political basis for economic success in China. World Polit 48(1):50–81
Murphy KM, Shleifer A, Vishny RW (1991) The allocation of talent: implications for growth. Q J Econ 106(2):503–530
Pejovich S (2003) Understanding the transaction costs of transition: it's the culture, stupid. Rev Aust Econ 16(4):347–361
Qian Y, Weingast BR (1997) Federalism as a commitment to perserving market incentives. J Econ Perspect 11(4):83–92
Rodrik D (1989) Promises, promises: credible policy reform via signaling. Econ J 99(397):756–772
Selgin G, White LH (2005) Credible currency: a constitutional perspective. Constit Polit Econ 16(1):71–83
Shleifer A (1997) Schumpeter lecture: government in transition. Eur Econ Rev 41:385–410
Stigler GJ (1966) The theory of price, 3rd Ed. New York: Macmillan
Storr VH (2013) Understanding the culture of markets. Routledge, New York
The World Bank (2017) World development indicators. World Bank Group, Washington, DC. https://openknowledge.worldbank.org/bitstream/handle/10986/26447/WDI-2017-web.pdf?sequence=1&isAllowed=y
Tullock G (1967) The welfare costs of tariffs, monopolies, and theft. West Econ J 5(3):224–232
Tullock G (1975) The transitional gains trap. Bell J Econ 6(2):671–678
Wagner RE (2016) Politics as a peculiar business: insights from a theory of entangled political economy. Edward Elgar, Northampton
Wagner RE, Runst P (2011) Choice, emergence, and constitutional process: a framework for positive analysis. J Inst Econ 7(1):131–145
Wagner RE, Yokoyama A (2013) Polycentrism, federalism, and liberty: a comparative systems perspective. J Public Finance Public Choice 31: 179–197
Weingast BR (1995) The economic role of political institutions: market-preserving federalism and economic development. J Law Econ Org 11(1):1–31
White LH (2010) The rule of law or the rule of central bankers? Cato J 30(3):451–463

Trial: Implicit Biases

Goran Dominioni[1] and Alessandro Romano[2]
[1]Rotterdam Institute of Law and Economics, Erasmus School of Law, Rotterdam, The Netherlands
[2]China-EU School of Law, China University of Political Science and Law, Beijing, China

Abstract

A large body of research in implicit social cognition indicates that implicit racial biases affect human decision-making. Building on these findings, legal scholars have explored how and under which circumstances implicit racial biases affect the functioning of trial systems. This entry reviews this literature. First, it provides an overview of the main results of, and methods employed in, studies on implicit racial biases. Second, it reviews the applications of these findings for the study of criminal and employment discrimination trials and the policies that could be implemented to reduce the effect of these biases on the decision-making of the relevant actors. It concludes with some suggestions for further research.

Definition

Implicit racial biases are shifts in judgment caused by automatic and/or unconscious attitudes/stereotypes held toward a racial group. In this context, automatic means that the bias occurs without any need for attention and that it is not easily controlled. Whereas unconscious means that introspection does not reveal the attitude/stereotype.

Introduction

In the last decades, studies in behavioral economics and psychology have highlighted the existence of various behavioral patterns that are inconsistent with the rational choice theory (Kahneman 2011 and references therein). Legal scholarship has

highlighted that many of these patterns are relevant for the study of law and policy-making (Jolls et al. 1998). Part of this literature, and especially in the context of trial settings and employment law, focuses on implicit biases (e.g., Jolls and Sunstein 2006; McAdams and Ulen 2009; Teichman and Zamir 2014). Although implicit biases do not relate only to race (but, for instance, also to gender), the wide majority of studies on this topic within behavioral law and economics focus on this particular issue. For this reason, the scope of this entry is restricted to implicit racial biases.

Scientific Basis and Research Methods

Two core concepts in the study of implicit biases are attitudes and stereotypes.

An attitude is a mental association between a racial group and an evaluative disposition (Greenwald and Krieger 2006). This evaluative disposition can be either positive or negative. For example, a person may hold either a positive or a negative attitude toward Asians. Instead, racial stereotypes are an association between a racial group and a positive or a negative characteristic (e.g., being lazy, good in math, or aggressive).

Research on implicit social cognition has studied the relationship between attitudes and stereotypes both at the implicit and the explicit level. This literature indicates that a person can hold an implicit attitude and a stereotype that point in opposite directions. For instance, it is possible to have a negative attitude toward Blacks and yet hold a positive stereotype toward them (e.g., Blacks are good in sports). In addition, existing evidence indicates that self-reported attitudes and stereotypes do not necessarily mirror implicit measures. As a consequence, it is possible for an individual to hold, and maybe act upon, an implicit attitude or stereotype that conflicts with the ones she consciously endorses (Greenwald and Krieger 2006).

Implicit attitudes and stereotypes are identified using implicit measurement procedures. Thanks to these techniques it is possible to identify attitudes and stereotypes that, for a variety of reasons, self-reporting may fail to detect (e.g., unawareness). Among the most commonly used techniques to measure implicit biases, there are affective priming, the implicit association test (IAT), and brain imaging.

In affective priming, subjects are exposed to two types of stimuli – a target and a prime – and they are asked to categorize the target as either positive or negative. To perform the task the only relevant stimulus is the target; therefore the prime should not affect the behavior of the subjects. For instance, in studies on implicit racial attitudes, subjects are first exposed to pictures of Black faces and White faces (the prime). This exposure occurs subliminally, meaning that pictures are shown for a very short time (e.g., 200 ms), so that they are processed only unconsciously. Subsequently, subjects are exposed to a series of words (the target) and are asked to categorize them (positive/negative). When priming with a racial stimulus that facilitates the categorization of the words as negative, the subject is said to hold a negative attitude toward the group (Fazio et al. 1995).

Another widely used measurement procedure in the implicit biases literature is the IAT (Greenwald et al. 1998). This procedure allows testing the relative strength of implicit attitudes and stereotypes via the measurement of response latencies in the categorization of stimuli into classes. In particular, in a racial IAT, part of the stimuli relates to racial groups (e.g., a name more commonly associated with a Black/White person), while the remaining stimuli relate to other concepts (e.g., good/bad). Subjects are repeatedly asked to categorize each stimulus (e.g., a Black name) as belonging to one of two dyads. Thus, for instance, subjects are first asked to categorize good/bad concepts and Black/White names as belonging to either the Black/good or the White/bad dyad. And, subsequently, the task is repeated with Black/bad and White/good dyads. Differences in response time between different combinations of dyads indicate that a certain racial group is more easily associated with a certain concept.

Another technique often employed in the implicit racial attitude domain is blood-oxygen level-dependent contrast imaging in functional magnetic resonance imaging (Fazio and Olson

2003). This physiological measure identifies variations in oxygen levels in the blood present in different parts of the brain, which are positively correlated with activations of these areas. Thus, for instance, exposure to Black faces has been shown to generate a greater activation of the amygdala (the amygdala is a part of the brain related to the processing of emotions). Therefore, this strand of research indicates that the exposure to different racial groups can trigger distinctive emotional states (Kubota et al. 2012).

Main Findings

The existing research indicates that implicit racial biases are pervasive in the White population of several Western countries. The largest database regarding the diffusion of implicit biases comes from the online platform Project Implicit, which allows visitors to take a racial IAT online. An analysis of the data gathered on this platform between the years 2000 and 2006 indicates that more than 65% of the people who took the racial IAT associated relatively more easily Blacks with bad. In addition, less than 15% of the participants (mainly non-White subjects) held relatively stronger associations between White and bad.

This data also suggests that Blacks are the only racial group that do not show strong pro-White associations (Nosek et al. 2007).

Research in implicit social cognition indicates that various factors mediate the formation of implicit biases. In particular, empirical evidence indicates that implicit biases can stem from past (often forgot) experiences, which may date back to early childhood. These experiences can be either direct or indirect. As indirect experiences – for instance, via the media – generally contribute to the formation of a negative perception of Blacks, this could explain the absence of a pro-Black bias among Black people (Greenwald and Krieger 2006).

A large part of the literature on implicit biases analyzes their impact on human behavior (both spontaneous and deliberate). Two meta-analyses on racial IAT studies find that measurements of implicit biases have a higher predictive validity than self-reported measures (Greenwald et al. 2009; Oswald et al. 2013).

A meta-analysis of racial priming studies reaches a similar conclusion (Cameron et al. 2012). Overall, this literature suggests that implicit biases can affect behavior, making them especially relevant for the study of discrimination in trial settings. In this regard, it is important to notice that even small effects size for a single decision can predict large discrimination at the societal level, especially when the bias affects many individuals or when it repeatedly affects one individual (Greenwald et al. 2015).

Applications in Trial Settings

Having introduced the background of the topic, let us now analyze how implicit biases relate to trial settings from a law and economics perspective. Broadly speaking, existing research on implicit biases in trial settings unfolds in two directions. First, studies in implicit social cognition made their ways into law reviews because of the impact that implicit biases have on trial actors' behavior. Second, implicit biases have been widely discussed among legal academics in relation to the epistemic value of implicit measurements in the context of evidence law. The second line of enquiry is sensu stricto not behavioral and is thus less relevant for the present discussion. In the following, we will therefore focus only on the behavioral side of the debate.

Implicit Biases and Trial Parties Decisions

Earlier law and economics literature depicts trial participants (judges, prosecutors, and lawyers) as rational agents that allocate resources following a rational calculation of the costs and benefits of their actions (McAdams and Ulen 2009). For instance, according to the efficient prosecutor model, the prosecutors maximize convictions (and attach different weights to different sentences) or seek deterrence maximization (Garoupa 2012). Later developments in the field questioned this simplistic description of trial participants and emphasized the role of institutional details and behavioral aspects of human decision-making.

Research on implicit biases is a subset of these developments.

With few exceptions, research on implicit biases in the courtroom has focused on criminal trials (Kang et al. 2012). In these settings, a figure that plays a major role in steering the unfolding of a trial is the prosecutor. Research suggests that implicit racial biases among prosecutors account for at least part of the disparities in incarceration rates between racial minorities and nonracial minorities in the US criminal law system (Smith and Levinson 2011; Kang et al. 2012). And indeed, implicit biases can be particularly effective in distorting human behavior when confronted with decisions that have the following characteristics: (i) allow for some discretion, (ii) have to be taken quickly, and (iii) do not provide accountability mechanisms. The decisions taken by prosecutors often present all these characteristics. A typical example is the decision on whether to charge a suspect or what crime to charge. In these cases, implicit stereotypes may influence prosecutors' decisions to the disadvantage of racial minorities. For instance, empirical evidence indicates that many individuals tend to have a stronger implicit association between Blacks and aggressive behaviors than between Whites and aggressive behaviors. Similarly, at the implicit level, Blacks are often more easily associated with guns than Whites. Smith and Levinson (2011) argue that these stereotypes may influence the decision of a prosecutor regarding whether to charge a crime both when the Black person is the alleged victim and when he is the alleged perpetrator. For example, following a shooting, prosecutors may have to decide whether to justify the act for self-defense. When the alleged perpetrator is Black and the victim is White, the bias might make it more plausible in the mind of the prosecutor that the harm caused by the Black was the product of an unjustified aggressive behavior. Instead, when the victim is a Black person the stereotype may lead the prosecutor to believe that his aggressive behavior may have justified the actions of the White person under investigation (Smith and Levinson 2011).

Another strand of literature focuses on defense attorneys. On this regard, the concern is twofold. On the one hand, implicit biases may induce a defense attorney to not defend a Black client. This, for instance, may occur when the attorney undervalues the probability of winning a case at trial on the basis of a biased evaluation of the available evidence. On the other hand, implicit biases may impair the performance of an attorney by undermining trust and communication with the client. Contextual factors such as work overload and degree of discretion may enhance the effect of implicit biases on attorneys' decisions. Incidentally, US public defenders often operate with insufficient resources to properly fulfill their tasks (Richardson and Goff 2013). Eisenberg and Johnson provide evidence of the pervasiveness of implicit biases among US defense attorneys (Eisenberg and Johnson 2004).

Most importantly, implicit biases may affect adjudicators' judgment and decision-making. And indeed, empirical evidence indicates that the judgment of both jury-eligible citizens and judges is affected by implicit biases (Hunt 2015; Kang et al. 2012). Regarding judges, Rachlinski and coauthors find that racial IAT scores among 120 US judges show similar patterns to those gathered with other samples. In particular, in line with the results obtained on the Project Implicit platform, they find that the large majority of White subjects held implicit attitudes that favor Whites over Blacks, while Black judges did not show strong preferences toward a particular racial group. The study further shows that when judges are not sufficiently motivated to control them, implicit biases affect their judicial decisions (Rachlinski et al. 2009).

Implicit biases can influence adjudicators' decisions from a number of perspectives, and in particular when judges have a relatively higher degree of discretion. Therefore, the influence of implicit biases might be more pronounced when the evidence presented by the parties is ambiguous (Kang et al. 2012). Hence, it is reasonable to conjecture that the effect of implicit biases might be more severe in civil cases, due to the lower burden of proof. In addition, implicit racial biases can affect memory recalls by leading adjudicators to remember facts in a more stereotypically consistent manner (Levinson 2007). In this regard, a

growing strand of literature is highlighting the existence of various implicit stereotypes that are of particular relevance in a criminal trial context. For instance, various studies suggest the existence of an implicit and bidirectional association between Blacks and crime (Hunt 2015). Similarly, Levinson and coauthors show the existence of a stronger implicit association between Blacks and guilt than Whites and guilt (Levinson et al. 2010). Importantly, IAT scores gathered with this measurement procedure predict evaluations of items evidence in a hypothetical trial. Along similar lines, in the context of death penalty judgments, implicit associations between racial groups and value of life have been shown to predict higher rates of death penalty punishments against Black defendants (Levinson et al. 2014). On a related note, implicit biases may partially account for the relatively higher rates of death penalty sentences inflicted to persons of color. For instance, implicit biases may ease finding the behavior of an offender as being heinous, atrocious, or cruel, which is one of the aggravating conditions under which death penalty can be inflicted under US law (Smith and Cohen 2012).

Outside the criminal law sphere, legal scholars have discussed implicit biases in the courtroom mainly with regard to employment discrimination lawsuits (Kang et al. 2012; Gertner and Hart 2012). In this context, the literature mostly focuses on the criteria that US judges are expected to adopt when deciding the dismissal of a case in its early stage. Here, the main concern regards the Iqbal standard, under which a judge should dismiss a case when the intent to discriminate the employee is not the most plausible explanation for the conduct of the employer.

In making this evaluation, the standard encourages judges to use their common sense. In this regard, legal scholars argue that, by asking judges to rely on common sense to make this decision, the law facilitates the influence of implicit biases on trial outcomes (Kang et al. 2012; Gertner and Hart 2012).

Debiasing and Insulating

Given the relevance of implicit attitudes and stereotypes, researchers are exploring how to decrease or eliminate their influence on trial actors' decision-making. From a general perspective, these interventions can act either on trial actors' general tendencies to be influenced by these biases or on the environment in which trial decisions are made.

One way to decrease the influence of implicit biases on the general decision-making of trial participants is via training (Rachlinski et al. 2009; Kang et al. 2012; Smith and Levinson 2011; Teichman and Zamir 2014). This training may aim at increasing awareness among trial actors about the existence of implicit biases. In turn, this may have the positive effect of decreasing actors' overconfidence in the objectivity of their own judgment while increasing their motivation to control the biases. Various studies in psychology confirm that motivation to control implicit biases can be effective in reducing their influence on behavior, and hence jurors should be advised to take the perspective of the out-group person (Kang et al. 2012). Rachlinski and coauthors further suggest that judges should take a racial IAT (Rachlinski et al. 2009). Beside increasing awareness, this may have the positive result of providing judges with a rough estimation of their biases and may therefore help them taking decisions regarding the training they may need as well as avoid overcorrections (i.e., shifting the bias against Whites).

Another strategy to reduce the impact of implicit biases is exposing trial actors' to counter-stereotypical information or increasing their contacts with members of racial minorities (Rachlinski et al. 2009; Kang et al. 2012; Smith and Levinson 2011). An obvious path is increasing racial diversification in the courtroom and in prosecutors' offices. However, this can only be effective as a long-term strategy. In the short term, debiasing can be attempted by introducing images of positive figures of minority members in judges' offices.

Alternatively, interventions could focus on the context in which decisions are made. For instance, trials could be structured so that trial agents are given sufficient time to make their decisions. And indeed, this could reduce their reliance on automatic processes and thus the influence of implicit

biases on their decision-making. Similarly, research indicates that conditions of high cognitive load (i.e., situations in which the cognitive activity imposed on working memory is high) ease the influence of implicit biases on judgment. Therefore, various authors suggest trial actors to adopt strategies that may reduce cognitive load when making decisions (Kang et al. 2012). Additional strategies include hiding racial information in the file that prosecutors use to decide whether to charge for a crime (Smith and Levinson 2011) and increasing jury diversity (Kang et al. 2012). In fact, research indicates that racially diverse juries tend to endure more thorough deliberations than all-White juries. In turn, this may reduce the reliance on automatic processes in decision-making.

Accountability can also play a role in reducing the effect of biases. For instance, it has been proposed that prosecution offices could gather data regarding the racial composition of individuals at each stage of the charging phase to provide useful feedback to prosecutors (Smith and Levinson 2011). Similarly, judges could keep track of their decisions, in order to spot potential systematic biases in their decision-making (Kang et al. 2012; Rachlinski et al. 2009). Another path to increase judges' accountability is increasing the depth of the scrutiny allotted to appellate courts in situations where the racial composition of the court of first instance fuels the suspect that the decision might be biased (Rachlinski et al. 2009).

Potential interventions to reduce the effect of implicit biases at trial are however not limited to public institutions, as also attorneys can implement strategies to help them overcome their biases. Public defenders could refer to objective standards for triage and checklists to evaluate their cases. These measures may help attorneys to impose clearer limits to their own discretion (Richardson and Goff 2013).

Future Directions

As discussed in this entry, studies in implicit social cognition can provide many insights on trial dynamics and remedies to contrast racial discrimination in the courtroom. Yet, this field of research is one of the latest developments of the already relatively young field of behavioral law and economics. Therefore, there is still a lot to learn on how these biases operate and how they influence the overall efficiency of the legal system. In the following, we highlight various possible pathways for future research.

First, it is still unclear how implicit biases affect the efficiency of legal systems. In particular, this strand of research can analyze how, to what extent and under which circumstances, implicit biases affect system costs related to criminal investigations, litigation, and adjudication. In addition, by influencing trial actors' behavior, implicit biases also affect the primary incentives provided by the law outside the courtroom (on a similar note see McAdams and Ulen (2009)). Last, as courts' accuracy is widely acknowledged as an essential element for the achievement of deterrence (Kaplow 1994), it is particularly relevant to study how implicit stereotypes affect accuracy in adjudication.

Second, the literature on implicit biases at trial mainly concentrates on Blacks (and Whites). However, the USA and other countries are becoming prismatic societies, and therefore more attention should be given to other racial and ethnic groups (e.g., Hispanics and Asians). Moreover, European scholarship and policymakers could benefit from research conducted on implicit discrimination against Arabs, Eastern Europeans, and Roma people. Along these lines, it might be warranted to conduct some studies on implicit stereotypes against Black people in Europe. In fact, as implicit biases are partially a product of culture, it might be interesting to study cross-cultural variations in implicit stereotypes.

Another pathway of research that remains largely unexplored regards implicit biases in non-criminal settings (Teichman and Zamir 2014). In particular, there is the need for further research in the fields of civil and administrative law. For instance, it has been argued that implicit biases may affect the application of the standard of proof in civil cases (Hunt 2015). Yet, these hypotheses await testing.

Future research can also contribute to a better understanding of debiasing mechanisms. As

shown above, the existing literature has started delving in this direction. Results reached so far are informative, but it is still not clear whether, and to what extent, these interventions are effective in real-life settings (Teichman and Zamir 2014). For instance, as discussed above, one of the main recommendations to decrease implicit biases among judges relates to increased diversity in judicial bodies. Yet, Rachlinski and coauthors found little differences in the pervasiveness of implicit biases between groups of judges coming from jurisdictions with great differences in racial diversity in the judiciary (Rachlinski et al. 2009). Further research could explore in more depth the conditions under which policies aimed at reducing implicit discrimination at trial are more likely to succeed.

Last, with few exceptions, most studies on implicit biases in the courtroom have been conducted with students. The field would benefit from having research conducted with professionals involved in trials. In particular, on the one hand, it would be interesting to gather data regarding the pervasiveness of implicit biases among trial agents. On the other hand, it would be important to analyze whether expertise affects the degree by which implicit biases impact behavior in trial settings.

Cross-References

- ▶ Behavioral Law and Economics
- ▶ Bounded Rationality
- ▶ Cognitive Law and Economics
- ▶ Judicial Decision-Making

References

Cameron CD, Brown-Iannuzzi JL, Payne BK (2012) Sequential priming measures of implicit social cognition a meta-analysis of associations with behavior and explicit attitudes. Personal Soc Psychol Rev 16(4):330–350

Eisenberg T, Johnson SL (2004) Implicit racial attitudes of death penalty lawyers. DePaul Law Rev 53(4): 1539–1556

Fazio RH, Olson MA (2003) Implicit measures in social cognition research: their meaning and use. Annu Rev Psychol 54:297–327

Fazio RH, Jackson JR, Dunton BC, Williams CJ (1995) Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? J Pers Soc Psychol 69(6):1013–1027

Garoupa NM (2012) The economics of prosecutors. In: Harel A, Hylton K (eds) Research handbook on the economics of criminal law. Edward Elgar, Cheltenham, pp 231–242

Gertner N, Hart M (2012) Employment law: implicit bias in employment discrimination litigation. In: Levinson JD, Smith RJ (eds) Implicit racial biases across the law. Cambridge University Press, New York, pp 80–94

Greenwald AG, Krieger LH (2006) Implicit bias: scientific foundations. Calif Law Rev 94(4):945–967

Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. J Pers Soc Psychol 74(6):1464–1480

Greenwald AG, Poehlman TA, Uhlmann EL, Banaji MR (2009) Understanding and using the implicit association test: III. Meta-analysis of predictive validity. J Pers Soc Psychol 97(1):17–41

Greenwald AG, Banaji MR, Nosek BA (2015) Statistically small effects of the implicit association test can have societally large effects. J Pers Soc Psychol 108(4):553–561

Hunt JS (2015) Race, ethnicity, and culture in jury decision making. Ann Rev Law Soc Sci 11:269–288

Jolls C, Sunstein CR (2006) The law of implicit bias. Calif Law Rev 94(4):969–996

Jolls C, Sunstein CR, Thaler R (1998) A behavioral approach to law and economics. Stanf Law Rev 50(5):1471–1550

Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York

Kang J, Bennett M, Carbado D, Casey P (2012) Implicit bias in the courtroom. UCLA Law Rev 59(5):1124–1187

Kaplow L (1994) The value of accuracy in adjudication: an economic analysis. J Leg Stud 23(1):307–401

Kubota JT, Banaji MR, Phelps EA (2012) The neuroscience of race. Nat Neurosci 15(7):940–948

Levinson JD (2007) Forgotten racial equality: implicit bias, decisionmaking, and misremembering. Duke Law J 57(2):345–424

Levinson JD, Cai H, Young D (2010) Guilty by implicit racial bias: the guilty/not guilty implicit association test. Ohio State Univ J Crim Law 8(1):187–208

Levinson JD, Smith RJ, Young DM (2014) Devaluing death: an empirical study of implicit racial bias on jury-eligible citizens in six death penalty states. NYU Law Rev 89(2):513–581

McAdams R, Ulen T (2009) Criminal behavioral law and economics. In: Garoupa N (ed) Criminal law and economics. Edward Elgar, Cheltenham, pp 403–436

Nosek BA, Hansen JJ, Devos T, Lindner NM, Ranganath KA, Smith CT, Olson KR, Chugh D, Greenwald AG, Banaji MR (2007) Pervasiveness and correlates of implicit attitudes and stereotypes. Eur Rev Soc Psychol 30(1):36–88

Oswald FL, Mitchell G, Blanton H, Jaccard J, Tetlock PE (2013) Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. J Pers Soc Psychol 105(2):171–192
Rachlinski JJ, Johnson SL, Wistrich AJ, Guthrie C (2009) Does unconscious racial bias affect trial judges? Notre Dame Law Rev 84(3):1195–1246
Richardson LS, Goff PA (2013) Implicit racial bias in public defender triage. Yale Law J 122(8): 2626–2650
Smith RJ, Cohen GB (2012) Capital punishment: choosing life or death (implicitly). In: Levinson JD, Smith RJ (eds) Implicit racial biases across the law. Cambridge University Press, New York, pp 229–243
Smith RJ, Levinson JD (2011) Impact of implicit racial bias on the exercise of prosecutorial discretion. Seattle Univ Law Rev 35(3):795–826
Teichman D, Zamir E (2014) Judicial Decisionmaking: a behavioral perspective. In: Zamir E, Teichman D (eds) The Oxford handbook of behavioral economics and the law. Oxford University Press, New York, pp 664–702

TRIPS Agreement

Daniel Gervais
Vanderbilt Intellectual Property Program, Vanderbilt University Law School, South Nashville, TN, USA

Abstract

The Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS Agreement) was negotiated between 1986 and 1994 during the Uruguay Round of the General Agreement on Tariffs and Trade (GATT), which led to the establishment of the World Trade Organization (WTO). The TRIPS Agreement sets minimum levels of several types of intellectual property (IP) protection, including copyright, trademarks, patents, industrial design, and trade secrets protection. Membership in the WTO includes an obligation to comply with the TRIPS Agreement. According to the WTO, the Agreement attempts to strike a balance between long-term social benefits to society of increased innovations and short-term costs to society from the lack of access to inventions (World Trade Organization (n.d.) Intellectual property: protection and enforcement. Retrieved from understanding the WTO: the agreements: http://wto.org/english/thewto_e/whatis_e/tif_e/agrm7_e.htm).

This entry considers this balance by looking at the two poles of intellectual property policy: providing incentives to increase innovation and optimizing access to inventions both for consumptive use and for potentially innovation-increasing experimentation. This entry also surveys the notion of *calibration*, the idea that every country or region should adapt its regulatory framework to reflect its own strengths and weaknesses in optimizing what one might refer to as its innovation policy. A calibration approach suggests that providing innovation incentives and optimizing access are not mutually exclusive objectives.

Part I: Incentivizing Innovation and Optimizing Access

The TRIPS Agreement has become the rope in the tug-of-war between utilitarianism and egalitarianism. Much of the Western intellectual property system seems to be based on a utilitarian theory regarding which IP rights should be given in order to incentivize the creation of new and potentially commercially valuable knowledge, referred to as "informational goods." This expression is meant to capture goods and related services that derive all or most of their value from the information they contain or embody.

As IP rights and protection standards increase, problems regarding access to existing informational goods arise. A maximalist utilitarian argument might be that there should be no right to access new informational goods because these goods would not have existed in the first place *but for* the existence of adequate IP rights. A more nuanced utilitarian approach might allow for some unlicensed use of certain informational goods, notably to maximize the continuing production of knowledge. Of course, in an optimal policy scenario, allowing for unlicensed uses of new goods should be done in a way that does not negatively affect incentives to innovate (Chandler and Sunder 2007). Egalitarianism, by contrast,

focuses mostly on distributional issues, notably access to new knowledge and goods.

Utilitarian and egalitarian objectives are not incompatible. In fact, incorporating egalitarian or distributive concerns into the utilitarian conversation can lead to the design of a system where both production and access are considered equally valid objectives. Unfortunately, discussions regarding the TRIPS Agreement have tended to focus either on the failure to produce sufficient incentives to innovate (i.e., the failure to supply an optimal supply of new informational goods) or on the failure to provide access to potentially life-changing and life-saving inventions. Both utilitarian and egalitarian perspectives are legitimate. From the utilitarian perspective, if new inventions or creations would not have existed *but for* the IP incentive, then negatively affecting incentives to innovate has adverse welfare impacts, including on economic development. This valid utilitarian argument can coexist with an equally valid egalitarian argument: in situations where markets fail to produce the optimal supply of the demanded informational goods due to anticompetitive practices or transaction costs, access may be improved by the use of appropriate exceptions to and limitations on IP rights, including compulsory licensing. In this scenario, access to these goods may be optimized without negatively impacting incentives to innovate in a significant way.

How does one strike the right balance of creating the incentives to enable or accelerate the development of new technologies and not restricting those technologies unduly when there is no real market to protect (e.g., in the case of a public good)? An economic analysis of IP law and policy makes it possible to dispel the normative fog of utilitarianism or egalitarianism, in which many legislative and judicial decisions are steeped. An economic lens offers novel ways of envisioning public actions and allows for the possibility of moving beyond inconsistencies and sticking points (Josselin and Marciano 2001). Striking the right balance requires the determination of what affects the value of an IP right, particularly what reduces that value.

Reduction of value may follow from granting (or even the threat of granting) a compulsory license or eliminating the right to exclude in a certain context by granting a limited legislative or court-made exception to allow for the free use of the protected creation or invention. While limiting exclusive rights may result in the reduction of the value of an IP right, it also allows third parties to understand, copy, and potentially improve on the subject of the IP right. The equilibrium between protecting the value of an IP right and optimizing the production of the protected informational good is difficult to achieve.

The Parameters of IP Policy Equilibrium

Normative debates based on the definition of property rights rarely provide a full answer to the policy question of balance outlined above. Intellectual property usually bears *aspects* of a property right, notably the right to exclude others, but that right is often much more limited than a full ownership right in a tangible object (Waldron 1988; Mackaay 1994). Consider the law of trademarks. It requires proof of a risk of confusion on the part of the potential purchaser before the right to exclude can be applied. No comparable requirement exists with respect to tangible objects. In the same vein, exceptions to the right to exclude such as fair use in copyright law, compulsory licensing of copyrights or patents, or experimental use of a patented invention without permission provide evidence of the varying nature of informational goods and of the relative nature of IP rights. Policy equations must reflect these differences.

The impossibility of determining the exact degree of overlap between property rights (in the classical sense of the term) and IP rights is not an outright impediment to striking an efficient policy balance using economic analysis. A list of objectives that is fairly persuasive can be drawn up: (1) the creation of sufficient incentives to innovate, (2) the efficient functioning of the markets for copyright works and patented inventions embedded in informational goods, and (3) the optimal (or at least reasonable) access to works and inventions. A fourth objective should be added, namely, the recognition (or attribution) of the work to the creator or inventor. An attribution right can be justified as useful for the author or inventor in generating goodwill but also as

providing users with information on the source of the good (e.g., name of the author, trademark of maker or supplier). A user can use this information to attribute trust value to the attributed source and thus possibly have decreased search costs in future transactions concerning the same good.

In order to achieve an equilibrium, IP policy equations must be solved in a way that meets the aims of all the interested parties – including creators, inventors, firms, investors, users, and reusers. The parameters of the solution allow for the perception of an adequate level of foreseeability for every group of stakeholders:

> [I]ntellectual property rights, like other property rights, set the parameters allowing investors to make a guess of the expected revenues. Only once these parameters are set will investments be undertaken. (Mackaay 1994)

Economic analysis leads policy-makers to set objectives, the achievement of which can be measured by concrete results rather than fulfillment of ideology.

Compulsory Licenses

An option to maximize access to and use of the invention or creation is to issue compulsory licenses. These licenses typically involve the payment of remuneration (determined to be adequate by competent authorities) and reporting requirements (amount made, used, exported (if any), etc.). Issuing compulsory licenses constitutes the establishment of a liability regime (instead of an exclusive right) (Reichman 1992).

Compulsory licensing is allowed under the TRIPS Agreement (World Trade Organization 2014). The debate regarding compulsory licenses, which has been on top of the agenda in international IP law and policy for decades, is centered on access to inventions but also on negative impacts on incentives to innovate. In this debate, the unfairness of the patent regime for indigent users who cannot afford new medicines is often emphasized; poor countries may only be able to procure new medicines at the expense of other urgent developmental priorities (i.e., at the expense of the production of other public goods).

Compulsory licenses may improve private parties' access to informational goods and help to achieve efficient outcomes when markets are working suboptimally in providing access, especially when access to the informational good concerned is considered necessary (e.g., pharmaceuticals or educational materials). Governing bodies issue compulsory licenses typically to allow the use of an informational good either to increase access or, second, in the hopes that third parties will make improvements or invent new products, which, in turn, will lead to competition in the market where the IP right owner is operating. The risk is the systemic weakening of incentives to innovate. If the price paid is considered too low, then financial incentives may be negatively impacted. This is also true if the instituted exception interferes with the markets operating in other territories. For example, one who purchases products at a lower cost due to a compulsory license should not export the products to another territory; such a practice may disrupt the market of the other territory. The holder of an exclusive right might also experience a reduction in the value of that IP right if a competitor were able to use the subject of the exclusive right without permission from the holder.

Compulsory licensing is subject to two additional caveats. First, knowledge that would be transferred with a *voluntary* licensing agreement may not accompany a *compulsory* licensing scenario. This knowledge may not be disclosed or otherwise "enabled" (meaning that a person with requisite knowledge can understand how to make and use the invention from the information disclosed in the patent) by the patent subject to the compulsory license. Second, a firm subjected to a compulsory license may withdraw, delay, or cancel other investments in the country. These are factors to bear in mind before issuing compulsory license. Their relative weight in the decision will vary case by case.

The Specific Case of Cultural Goods

There are economically relevant cultural aspects to the creation of incentives to innovate in the literary and artistic fields. Problems of incentives and access can impact the creation and viability of industries that produce and market cultural informational goods. Additionally, IP rules can negatively impact access and use of such goods.

Incentives to innovate are linked to the size of potential markets for new goods, which in gross economic terms would factor in the number of prospective users and their financial ability to purchase "cultural goods," which are a subset of informational goods more directly linked to cultural and entertainment industries, such as films, music, and books in various formats. This explains the importance of English – and specifically of the North American market – in this context. Because many buyers in those regions can afford to purchase many cultural goods, they create a solid market for English-language informational goods and thus an incentive to produce such goods. The role of English as a second language in many other parts of the world reinforces this feature.

When goods are produced for markets where prices are comparatively high, they may not be made available at lower prices in other markets, in part because price discrimination may lead to exports from lower-priced markets to higher-priced ones. Such exports may be unauthorized by the copyright holder but still legal in countries that allow "international exhaustion," namely, the ability to import in their territory goods legally produced in other countries. The TRIPS Agreement contains no mandatory rule on exhaustion.

In 1971, an appendix was added to the main international copyright instrument, the Berne Convention. This appendix allows compulsory licenses to be issued for the translation and reproduction of foreign books in developing countries in order to increase access to knowledge. Developing countries, particularly their education systems, may not have the economic ability to purchase cultural goods in sufficient quantity to validate the incentive to innovate protected by the IP system. Because some developing countries do not have sufficient buying power (compared to more industrialized markets), they would not contribute to the monetary award that the IP right entitles the creator of the work to receive. The ability to issue compulsory licenses for translation has in fact been one of the most contentious issues since the early days of the Berne Convention.

Part II: Calibration Within TRIPS Parameters

Changes to the TRIPS Agreement and the adoption and implementation of new wide-ranging multilateral treaties are unlikely and perhaps decades away. In fact, it took over a decade to pass Article 31*bis*, the only amendment to the TRIPS Agreement thus far. Therefore, at this juncture and for the predictable future, the calibration of IP policy will remain mostly a matter of domestic policy-making and regional trade agreements. Calibration will most likely involve using flexible provisions contained in TRIPS (i.e., the specific exceptions and limitations) or interpreting the Agreement to achieve IP policy goals. For example, courts may interpret the Agreement to include fair use/fair dealing exceptions to copyright rights and setting of appropriate limits on patentable subject matter.

Structurally, calibration is not a rejection of the harmonization mentioned in the opening paragraph. The calibration approach recognizes that rules will vary to a certain degree due to differences among regions, countries, and industries. This approach suggests that variations within parameters set by TRIPS are *desirable*. Calibration suggests that, by developing a comprehensive IP strategy focused on innovation and welfare improvements, a country can limit the negative impact of transitioning to more rigorous IP protection and increase its chances of reaping the benefits thereof, including technology-related foreign direct investment (FDI) and growing domestic Web-based, pharmaceutical, and other technology-intensive industries.

The Calibration Process

How does one design a calibrated implementation of TRIPS as part of an innovation optimization strategy? One can begin by eliminating what is unlikely to work. For example, there are probably no definitive answers given to questions such as whether the optimal term for a patent is 12, 18, or 22 years. All one can say is that, beyond a certain point, patent protection *decreases* innovation (Gallini 1992). Empirical studies have verified that in countries without the necessary technology-

absorptive capacity, increasing patent protection for pharmaceuticals produces little, if any, innovation outcomes (La Forgia et al. 2009). Similarly, in the copyright realm, there are no definitive answers to how long IP protection should last. Is the optimal copyright term 14 years from publication or the life of the author plus zero, 50, or 70 years? (Sprigman 2004). What combination of rights and exceptions would better achieve the policy goals than rules in place now? Is using a *single* term of protection part of the problem? One could argue that for a certain invention (protected by patent) or creation (protected by copyright), one specific term is optimal while a different term is more appropriate for another invention or creation (Stanley 2003). Finding the optimal point of IP protection is difficult because each informational good might require a different level of protection. Thus, horizontal rules (such as a single term of protection or bundle of rights) are approximations at best. These policy discussions can take the form of continuous imprecise balancing acts, making it difficult to reach a calibration target (Falvey et al. 2004).

Instead of engaging in *systemic* analyses to issue bright-line rules (e.g., patent protection lasts 20 years), policy-makers could try the other extreme, that is, a full case-by-case approach. Such an approach would factor in the exact value added of *each* informational good. Value determination would then depend on measuring the exact size of the inventive step involved in the invention concerned. This is related to the issue of measuring *patent quality*, and various metrics have been proposed, including by the Organisation for Economic Co-operation and Development ("OECD"), by considering how many times a patent is cited in later patents (Organisation for Economic Co-operation and Development 2011). Such measurements often emphasize the need to move policy goals away from purely quantitative patent metrics (number of applications or grants) and focus on the quality of the information disclosed.

One could add to this equation the degree of competition in the industrial or economic sector impacted by the invention and, relatedly, the number of dominant players by market share (Hollis 2004). Even if such a case-by-case approach did not encompass high transaction costs, experts could only estimate the future utility of the invention. In terms of predictability, time, and transition/protection costs, a system of copyright or patent with a single set of rights and one term of protection may therefore be better than case-by-case analysis. While perhaps theoretically less attractive, such a system is simple to understand and easy to administer.

Does this mean that one is stuck at the abstract level without any hope of achieving relative analytical clarity? Here, TRIPS only offers normative guideposts and guarantees minimum levels of protection. In implementing TRIPS, each country or region should calibrate its IP policy recognizing its own unique characteristics and needs.

Calibration by Region and Industry

Both the systemic and case-by-case approaches are suboptimal for the aforementioned reasons. A more realistic approach for policy-makers trying to effectuate calibration revolves around the level of rights and exceptions that, *within the range of TRIPS-compatible implementations*, is most likely to achieve the policy goal of maximizing innovation while minimizing negative welfare impacts. Because the TRIPS Agreement only "harmonized" national laws to a degree, it left space for calibration (Reichman 2000). One should use a combination of human *and* economic development factors in order to craft a set of objectives to help determine the best level of rights and exceptions. These objectives will be amenable at least to the kind of metrics (e.g., the United Nations Human Development Index) that evidence-based policy-making calls for. Development is "a pseudonym for a complex network of benefits associated with economic growth and human social capital" (Okedjii 2009). It is the sum of the changes in social patterns and norms through which production devices couple with the population: the latter acquires the capacity to utilize the former in order to achieve what is considered to be a satisfactory growth rate, and the production devices supply products that serve the population instead of being "alien" to the population. This dialectic of production devices

and population is the essence of development (Perroux 1988).

Establishing broad developmental goals is not the same as implementing them. Each country implementing TRIPS must recognize how it compares to others in the region or countries elsewhere at a similar level of development.

Differences Among Countries

Developing countries are not all identical; in fact, they are far from identical. Developing countries can be grouped in various ways. Leaving aside the least developed countries (LDCs), for which TRIPS obligations have been suspended, one could distinguish developing countries via a "net outcomes" approach. Countries in which innovation benefits outweigh additional rent extraction can be grouped together, and countries in which additional rent extraction outweigh innovation benefits can be grouped together. Professor Llew Gibbons offers a promising taxonomy of developing countries (Gibbons 2011). He grouped these countries according to three stages of development:

Stage One
- Foreign direct investment is rare and usually limited to specialized sectors – often relating to the exploitation of natural resources or developing franchise service industries like a major international brand bottling company.
- There exists unskilled, cheap labor.
- Foreign businesses create the necessary infrastructure and invest in human capital.
- The developing country is investing in the training of skilled workers and junior managers. Successfully developing a skilled workforce is a prerequisite to entering stage two.

Stage Two
- The economy is now able to absorb technology, to imitate technology at some level, and to contribute minor improvements.
- There exists a well-educated workforce adapted to absorb new technology from other countries and then incorporate the technology into the domestic economy.
- Domestic research efforts are primarily facilitative or associated with technology transfer. Focus shifts gradually to efforts on more innovative projects.

Stage Three
- The newly industrialized country produces its own intellectual property.
- The country is very selective as to which IP rights it zealously protects.

This progression from imitation to absorption to innovation tracks the path proposed by innovation theorists, i.e., progression from imitation to adaptation to true global innovation. Countries have indeed followed the progression described; along the pathway of development, the countries have gradually created more IP rights and made better use of existing IP rights. As countries build innovation-focused industries, they generally develop more sophisticated and nuanced IP policies (Gervais 2005). Simply put, they play the IP game better.

Differences Among Industries

The TRIPS Agreement put all industries on the same footing. Yet, treating all industries as equally sensitive to IP protection is incorrect and results in a number of unintended consequences. First, there is a relatively short list of industries generally considered to be highly patent sensitive. The list includes industries consisting of certain chemical companies, producers of laboratory instruments, and makers of steel-mill products (Cheng 2012; Maskus 2000; Gervais 2005). This list is partially derived from Mansfield's studies of the field, in which he found that while 90% of pharmaceutical innovations were dependent on patent protection, only 20% of electronics and machinery innovations were patent dependent (Cheng 2012; Mansfield 1986). Beyond the general consensus on the patent dependency in these aforementioned fields, controversy concerning the role and impact of patents on innovation quickly emerges. Do the benefits from patents to computer software and online commerce outweigh the costs of obtaining, enforcing, and defending against patent infringement claims, especially those brought by

nonpracticing entities? (Mullin 2013). This question is front and center in a number of both developed and developing countries (Shrestha 2010). Proposals to recognize differences in industries have been made in the United States. For example, some scholars suggest that if the legislator is unwilling to separate industries by applying different standards, then perhaps courts should (Burk and Lemley 2010). In a developing country, implementing a *single* patent policy for *all* industries – as is facially required by TRIPS – may thus be structurally suboptimal.

While the TRIPS Agreement identifies uniform patentability criteria, the question whether it makes sense to treat all industries the same should be on any comprehensive innovation agenda. A form of "discrimination" based on the nature of the industry concerned would be justified if one could develop a proper metric to measure whether and how innovation outcomes are achieved.

Conclusion

As policy-makers around the world navigate deeper in the calibration waters of TRIPS implementation and seek to optimize national innovation strategy within the boundaries ascribed by TRIPS, they must attempt to avoid the negative consequences of increasing IP protection, from patent trolls, and the prevention of access to essential medicines by patents to the chilling of free expression and prevention of fair uses by copyrights and trademarks. While avoiding the Scylla of excessive IP, one must also not run to the Charybdis of insufficient IP. This is the inevitable and desirable process of calibration that an economic analysis of IP law can both inform and guide.

Cross-References

- ▶ Copyright
- ▶ Intellectual Property: Economic Justification
- ▶ Trade Secrets Law
- ▶ Trademarks and the Economic Dimensions of Trademark Law in Europe and Beyond
- ▶ WTO: Procedural Rules

References

Burk D, Lemley M (2010) The patent crisis and how the courts can solve it. Syracuse Sci Tech Law Rep 23:1–22

Chandler A, Sunder M (2007) Is Nozick kicking Rawls ass?: intellectual property and social justice. UC Davis Law Rev 40:563

Cheng T (2012) A developmental approach to the patent-antitrust interface. Northwest J Int Law Bus 33:1–79

Falvey R, Foster N, Greenaway D (2004) Intellectual property rights and economic growth. Internationalisation of economic policy research paper no 2004/12, 2. Retrieved from http://ssrn.com/abstract=715982

Gallini N (1992) Patent policy and costly imitation. Rand J Econ 23(1):52–63

Gervais D (2005) Intellectual property and development: the state of play. Fordham Law Rev 74(74):505–535

Gibbons L (2011) Do as I say (not as I did): putative intellectual property lessons for emerging economies from the not so long past of the developed nations. SMU Law Rev 64(3):923–973

Hollis A (2004) An efficient reward system for pharmaceutical innovation. Retrieved from http://www.who.int/intellectualproperty/news/Submission-Hollis6-Oct.pdf

Josselin J-M, Marciano A (2001) L'analyse economique du droit et le renouvellement de l'econonmie politique des choix publics. Economie Publique/Pub Econ 7:6

Mackaay E (1994) Legal hybrids: beyond property and monopoly? Columbia Law Rev 94:2630

Mansfield E (1986) Patents and innovation: an empirical study. Manag Sci 32(2):173–181

Maskus K (2000) Lessons from studying the international economics of intellectual property rights. Vanderbilt Law Rev 53(6):2219–2239

Mullin J (2013) In historic vote, New Zealand bans software patents: patent claims can't cover computer programs 'as such'. Retrieved from Ars Technica: http://arstechnica.com/tech-policy/2013/08/in-historic-vote-new-zealand-bans-software-patents

Okedjii R (2009) History lessons for the WIPO development agenda. In: Netanel W (ed) Intellectual property and developing countries. Oxford University Press, Oxford, pp 137–162

Organisation for Economic Co-operation and Development (2011) Competing in the global economy, technology: quality in OECD science, technology and industry scoreboard. Retrieved from oecd.org

Perroux F (1988) The pole of development's new place in a general theory of economic activity. In: Higgins B, Savoie D (eds) Regional economic development: essays in honour of Francois Perroux. Unwin Hyman, Boston, pp 48–76

Reichman J (1992) Legal hybrids between the patent and copyright paradigms. In: Altes WK (ed) Information law towards the 21st century. Kluwer Law, Deventer, pp 325–341
Reichman J (2000) The trips agreement comes of age: conflicts or cooperation with the developing countries. Case West Reserve J Int Law 32:441
Shrestha S (2010) Trolls or market-maker? An empirical analysis of nonpracticing entities. Columbia Law Rev 110:114–160
Sprigman C (2004) Reform(aliz)ing copyright. Stanf Law Rev 57:552
Stanley C (2003) A dangerous step toward the over protection of intellectual property: rethinking Edler v. Ashcroft. Hamline Law Rev 679:694–695
Waldron J (1988) The right to private property. Clarendon, Oxford
World Trade Organization (2014) Members accepting amendment of the TRIPS Agreement. Retrieved from intellectual property: trips and public health: http://www.wto.org/english/tratop_e/trips_e/amendment_e.htm
World Trade Organization (n.d.) Intellectual property: protection and enforcement. Retrieved from understanding the WTO: the agreements: http://wto.org/english/thewto_e/whatis_e/tif_e/agrm7_e.htm

Further Reading

La Forgia F, Osenigo L, Montobbio F (2009) IPRs and technological development in pharmaceuticals: who is patenting what in Brazil after TRIPS. In: Netanel W (ed) The development agenda: global intellectual property and developing countries. Oxford University Press, Oxford, pp 293–319
Nachbar T (2004) Intellectual property and constitutional norms. Columbia Law Rev 104:338–339
Prud'homme D (2012) Dulling the cutting-edge: how patent-related policies and practices hamper innovation in China. Beijing, European Union Chamber of Commerce in China
Ragavan S (2012) Patent and trade disparities in developing countries. Oxford University Press, Oxford

Type-I and Type-II Errors

Matteo Rizzolli
LUMSA University, Rome, Italy

Abstract

Adjudicative procedures meant at establishing truth about facts on defendants' behavior are naturally prone to errors: defendants can be found guilty/liable when they truly were not (type-I errors) or they can be acquitted when they should have been convicted (type-II errors). These errors alter the incentives of defendants to comply with norms. We review the literature with a particular focus on type-I errors.

Introduction

The word *adjudication* has its Latin roots into the words *jus* (right, justice) and *dicere* (to say). All organizations set goals and have adjudicative procedures to "establish the truth" about members' compliance. Performance appraisal committees decide whether employees earn rewards within business organizations; teachers must assess students' progress in learning; courts adjudicate whether citizens have committed crimes; disciplinary committees within sport leagues and professional organizations as well as religious tribunals assess whether members' conduct has been conforming. Adjudicative procedures must evaluate and reward something they cannot directly observe – being it effort, intention, or act – and this makes them obviously prone to errors. Although our framing will be mostly applied to criminal (See aslo ▶ "Criminal Sanctions and Deterrence" and ▶ "Crime, Incentive to"), administrative, and civil courts, most of the results here presented apply to any generic adjudicative procedure. We thus consider the general case of an adjudicative authority who (i) must assess whether the observed behavior of an individual *conforms* or *deviates* from the prescribed behavior and (ii) must incentivize or sanction such behavior accordingly. In judging behavior, errors inevitably arise, and they generally undermine individuals' incentives. These errors take mainly two forms: (i) the adjudicative authority may assess non-compliance when in fact the subject is duly complying and (ii) the adjudicative authority may assess compliance when in fact the subject is deviating. Individual's compliance with the prescribed behavior can be interpreted as the null hypothesis, so that the adjudicative authority can both incorrectly reject the null and sanction a complying subject (a type-I error) and incorrectly accept the null and exculpate an undeserving subject (type-II error). In the context of crime

deterrence, type-I errors amount to wrongful convictions of innocents. We model the relation between type-I and type-II errors below within a standard optimal deterrence framework. Finally, we discuss the empirical relevance of type-I errors.

Basic Setup

Let y_0 be the initial endowment equal for all agents and b the benefits from deviating from the prescribed behavior (e.g., committing crime). b is distributed among the agents with a generic distribution $z(b)$ and a cumulative $Z(b)$ with support $[0\bar{B}]$. Let also h be the harm/externality generated by each individual's deviation (each individual takes this decision only once). For the sake of simplicity, all individuals are audited and brought in front of an adjudicative authority. The authority observes the amount of inculpatory evidence e that is produced against a defendant, and if this overcomes a certain threshold, $\tilde{e}$ then the authority imposes a monetary sanction s. For the sake of simplicity, we also assume that there is no welfare-improving deviation as in Becker seminal paper on crime (See also ▶"Crime and Punishment (Becker 1968)" by Becker 1968 and also to "Becker, Gary S.") (this would be a crime for which $b > h$) and that monetary sanctions are transferred from the defendant to society. Furthermore, in the function of social costs, we do not consider the private benefits from crime but only its social costs.

Therefore let e have a frequency distribution of $i(e)$ for the conforming defendant (innocent) and of $g(e)$ for the deviating defendant (guilty). Let $I(e)$ and $G(e)$ be the cumulative distributions of $i(e)$ and $g(e)$, respectively, and note that $I(\tilde{e})$ and $G(\tilde{e})$ are the probabilities of being acquitted for the complying and for the deviating defendant, respectively, given the evidence threshold $\tilde{e}$. To keep notation compact, we will often use I and G for $I(\tilde{e})$ and $G(\tilde{e})$, respectively.

The evidence is stochastically distributed, albeit in general more incriminating evidence is available against deviating defendants than against complying ones. First-order stochastic dominance is assumed $I(e) > G(e)\ \forall\ e\ \in]0, e_{\max}[$. Without FOSD evidence would be produced randomly for the complying and the deviating alike, and therefore the whole criminal procedure would be pointless. Note also that G is the probability of type-II error and $1 - I$ is the probability of type-I error. Let us also define $\Delta(\tilde{e}) = I - G$ as the *accuracy* of the adjudicative procedure; Δ represents the ability of the procedure to distinguish complying from deviating defendants.

For our purpose, we assume the social planner optimizes deterrence only by affecting the threshold $\tilde{e}$ which in turn determines the error's trade-offs: for instance, an increase in $\tilde{e}$ generates both an increase in the number of wrongful acquittals G and a decrease in the number of wrongful convictions $1 - I$.

The risk-neutral individual does not deviate as long as the returns from deviating behavior are smaller than the expected returns of conforming. Since b varies across individuals, there exists a level of $\tilde{b}$ for which the individual is indifferent between conforming and not, and this determines the proportion of the population $Z(\tilde{b})$ who conforms.

Social welfare is thus $(1 - Z(\tilde{b}))h$: the social costs of harm caused by those defendants who deviate. On the other hand, the social planner only acts on the threshold $\tilde{e}$ which implicitly defines the trade-off between type-I and type-II errors. The link between the social planner's choice of the evidentiary standard $\tilde{e}$ which in turn determines the error's trade-off and the defendant's choice of conformity determined in $\tilde{b}$ are the ingredients to understand the role of adjudication in deterrence.

Let us begin by assuming agents to be risk-neutral utility maximizers. The returns from conforming are $E\pi_I = y_0 - (1 - I)s$, while the returns from deviating are $E\pi_G = y_0 + b - (1 - G)s$. All defendants for which $E\pi_i \geq E\pi_g$ will conform and therefore the threshold level of b which implicitly defines the conforming population is

$$\tilde{b}_{rn} = (1 - (1 - I) - G)s \qquad (1)$$

By looking at Eq. 1, we can single out the typical "deterrence effect" as $\tilde{b}$ increases both with the magnitude of the sanction $(\uparrow s \Rightarrow \uparrow \tilde{b})$

T

and via an increase in the detection probability for the deviating defendants which in this model corresponds to a decrease in the probability of type-II errors $\left(\downarrow G \Rightarrow \uparrow \tilde{b}\right)$. Furthermore, a "compliance effect" of type-I errors can be seen: s $\tilde{b}$ increases when the probability of being punished decreases for conforming defendants $\left(\uparrow I \Rightarrow \uparrow \tilde{b}\right)$. Also a "screening effect" can be established: the higher is the accuracy Δ, the better the procedure can discriminate between conforming and non-conforming behaviors and the greater the advantages of staying conforming ($\uparrow \Delta \Rightarrow \uparrow \tilde{b}$). Finally, by simple inspection of Eq. 1, it is evident that marginal change in either $1 - I$ or G determines an equal decrease of $\tilde{b}$ as $\frac{\partial b}{\partial(1-I)} = \frac{\partial b}{\partial G} = s$. Under risk neutrality, type-I errors $(1 - I)$ and type-II errors (G) have the same negative impact on the defendant's incentive to comply. This is because on one hand type-II errors undermine compliance inasmuch as they decrease the probability of nonconforming defendants being finally sanctioned. On the other hand, type-I errors increase the opportunity costs of conforming relative to deviating.

Now that the threshold level $\tilde{b}$ is defined, the social welfare can be computed and derived with respect to the evidence threshold:

$$\begin{aligned} \frac{\partial SW}{\partial e} &= \partial\left(1 - Z\left(\tilde{b}_{rn}\right)\right)h \\ &= -z\left(\tilde{b}_{rn}\right)(i(\tilde{e}) - g(\tilde{e}))sh \end{aligned} \tag{2}$$

Let $\tilde{e}_{\text{neutral}}$ be implicitly defined by $i(\tilde{e}) = g(\tilde{e})$. By inspection of Eq. 2, the optimal evidence threshold $\tilde{e}$ that minimizes social costs is $\tilde{e}_{\text{neutral}}$. In fact accuracy reaches its maximum level when the social planner chooses $\tilde{e}_{\text{neutral}}$ so that the distance between the two cumulative functions is maximized. If the social planner chooses a higher evidence threshold $\tilde{e}_{\text{pro-def endant}} > \tilde{e}_{\text{neutral}}$, then the error trade-off tilts in favor of the defendant as the probabilities of both correct and wrongful acquittals $-I$ and G, respectively, increase. $\tilde{e} > \tilde{e}_{\text{neutral}}$ necessarily also implies $g(\tilde{e}) > i(\tilde{e})$ by definition of the frequency distribution of $i(e)$ and $g(e)$. Notice that for levels of $\tilde{e} > \tilde{e}_{\text{neutral}}$, G grows faster than I and therefore accuracy cannot be maximal.

Evidence Thresholds, Standard of Evidence, and Error Ratios

While our analysis focuses on the evidentiary threshold $\tilde{e}$ that determines the probabilities of both type-I and type-II errors, there are other two common concepts that concern adjudication and that must be put in relation with our analysis.

The first one is the **standard of evidence**: it is generally understood as the level of certainty the adjudicative authority must reach in order to establish guilt in a criminal proceeding (or liability in civil one). Among the most common standards of proof used in different adjudicative procedures, there are the *preponderance of evidence* (*poe*) standard, the *clear and convincing evidence* (*cace*) standard, and the *beyond any reasonable doubt* (*bard*) standard. Although giving probabilistic interpretations of these standards of proof is very controversial (see Kaplow 2012, footnote 76 for a discussion), they are commonly understood to roughly coincide with the 50%, 75%, and 95% thresholds, respectively. Under *poe* (or *cace* or *bard*), the adjudicative authority must answer to the question of whether, given the evidence available, the likelihood that the defendant has deviated is larger than 50% (or 75% or 95% depending on the standard applied). These probabilities must be understood as Bayesian posterior probabilities of having deviated, and these are functions – following the Bayes' rule – of the likelihood of the signal given by the densities i and g of the evidentiary threshold $\tilde{e}$ and on the prior probability of being brought in front of the adjudicative authority. The probability that a defendant has deviated or not also depends on the base rates of the two actions, $\left(1 - Z\left(\tilde{b}\right)\right)$ and $\left(Z\left(\tilde{b}\right)\right)$, respectively. The $\tilde{b}$ are determined endogenously by defendants' decisions and ultimately depend on the evidence threshold $\tilde{e}$. Therefore in order to identify the proper threshold $\tilde{e}$ – in case the *poe* standard applies – one should ask what value of $\tilde{e}$ implicitly solves the equation $g(\tilde{e}) \cdot \left(1 - Z\left(\tilde{b}\right)\right) = i(\tilde{e}) \cdot \left(Z\left(\tilde{b}\right)\right)$. If the adjudicative authority needs to apply the *cace* or *bard* standard, one could simply multiply by either 3 or 19 the right side of the previous equation. As Lando

(2002) and Kaplow (2012) point out, the two notions – the one based on the optimal *evidence threshold* and the one based on the *standard of evidence* – are strikingly different. To begin with, the optimal evidence threshold is derived from welfare analysis and seeks to find the level of $\tilde{e}$ that maximizes social welfare. By contrast, within the *standard of evidence* framework, $\tilde{e}$ is derived by asking under what circumstances would the probability that the defendant before the adjudicative authority has actually deviated be 50% (or 75% or 95% or other conventional probabilities). In fact the optimal *evidence threshold* could be associated with any probabilistic *standard of evidence* whatsoever.

Another approach focuses on the **ratio of errors** and expresses the pro-defendant bias of adjudicative procedures in terms of error ratios. There seems to be something specific about type-I errors in the context of crime: scholars and rule makers across time and societies advocated a specific attention to the avoidance of type-I errors even at the expense of many type-II errors; arguably the most famous statement in this respect is the one of William Blackstone (1769) recommending that *it is better that ten guilty persons escape than that one innocent suffer.* Dekay (1996) systematizes the relation between the standards of evidence and the error ratios. We can interpret these as ratios of errors' frequencies where the frequency of erroneous acquittals is the conditional probability that a truly deviating defendant is acquitted (type-II error) multiplied by the base rate of the action $\left(1 - Z\left(\tilde{b}\right)\right)$, while the frequency of erroneous conviction is the conditional probability that a truly complying defendant is convicted (type-I) multiplied by the base rate $\left(Z\left(\tilde{b}\right)\right)$. So the type-I error ratio (sometimes also called the Blackstone's error ratio) is defined as $\frac{G\cdot\left(1-Z\left(\tilde{b}\right)\right)}{(1-I)\cdot Z\left(\tilde{b}\right)}$.

All else being equal, higher standards of evidence that affect the trade-off between G and I do imply higher Blackstone-like error ratios. However, it should be noticed that the optimal *evidence threshold* could be associated with many different error ratios depending on the base rates.

Risk and Loss Aversion

Subjects are known to be generally risk-averse in their utility of income. We thus introduce risk aversion in the measure of the monetary gains from crime b following Rizzolli and Stanca (2012). If b are monetary gains for which utility $U(.)$ can be derived, then the expected utility of complying is $IU(y_0) + (1 - I) \cdot U(y_0 - s)$, while the expected utility of deviating is $GU(y_0 + b) + (1 - G) \cdot U(y_0 + b - s)$. The threshold level of $\tilde{b}_{eu}$ that triggers a defendant to deviate is implicitly defined by

$$\begin{aligned} &I[U(y_0) - U(y_0 - s)] \\ &- G[U(y_0 + b) - U(y_0 + b - s)] \\ &\geq U(y_0 + b - s) - U(y_0 - s) \end{aligned} \tag{3}$$

Equation 3 shows that when there is an increase in either of the errors (increase in G or decrease in i) on the left-hand side of the equation, defendants find deviation convenient for lower levels of b (on the right-hand side). However, given the concavity of the utility function, the negative impact of type-I errors $(1 - I)$ on the threshold level $\tilde{b}_{eu}$ and thus on social welfare is stronger than that of type-II errors (G). To see why, note that $U(y_0) - U(y_0 - s) > U(y_0 + b) - U(y_0 + b - s)$. In order to maintain the same level of deterrence, a given percentage increase of $1 - I$ must be compensated by a smaller percentage decrease of G. Therefore, assuming risk aversion, type-I errors $(1 - I)$ create more disutility and thus induce more deviation than comparable type-II errors (G); therefore, social costs are minimized for a $\tilde{e}^* > \tilde{e}_{\text{neutral}}$. The opposite result holds if we instead assume risk-seeking behavior.

Another interesting extension concerns the introduction of loss aversion: a departure from the expected utility framework that has been incorporated in models such as the cumulative prospect theory (Dhami and al Nowaihi 2013). These models build on the empirical observation that people tend to think of possible outcomes of a choice under uncertainty relative to a certain reference point and tend to prefer the avoidance of losses (outcomes below the reference point) than the acquisition of comparable gains (outcomes above the reference point). Incorporating

reference-dependent preferences and loss aversion in the model is not trivial (see Nicita and Rizzolli 2014); however, the intuition and the results are quite simple: type-I errors always imply a potential loss relative to the status quo, while this is not necessarily true for type-II errors. To conclude, in presence of loss aversion, type-I errors $(1 - I)$ represent a net loss and impact the defendant value function more than comparable type-II errors (G); therefore, social costs are minimized for a $\tilde{e}^* > \tilde{e}_{\text{neutral}}$.

Cost of Sanctions

So far we have assumed that the sanction s is monetary and that it implies – once imposed – a costless transfer from the defendant to the society. However, the imposition of sanctions implies both private costs of punishment to defendants and to society as well. Nonmonetary sanctions are a social cost (Shavell 1987) as their imposition implies a disutility for the defendant that is not transferred to society. Furthermore, all sanctions – including monetary fines – must be administered and therefore imply a social cost (Polinsky and Shavell 1992).

Define c as the total cost (both to the defendant and to the society) of imposing a sanction. The social welfare function (assuming risk neutrality) should be rewritten as the following:

$$SW = \left[1 - Z\left(\tilde{b}\right)\right]h + \left[1 - Z\left(\tilde{b}\right)\right](1 - G)c + Z\left(\tilde{b}\right)(1 - I)c \tag{4}$$

The first term of Eq. 4 represents the harm/externality of deviating, as before. The second term represents the expected total costs of imposing sanctions on deviating defendants, and the third term represents the expected total costs of punishing complying defendant (type-I errors). The problem lies in defining the optimal $\tilde{e}$ that minimizes the expected total costs from crime, including the costs of punishment. As before, the first term is minimized for $\tilde{e} = \tilde{e}_{\text{neutral}}$. However, the second and third terms are minimized for $\tilde{e} \to \infty$. In fact for an evidence threshold $\tilde{e}$ arbitrarily high, the probability of correctly imposing a sanction on a deviating defendant $(1 - G)$ or erroneously imposing a sanction on a complying defendant $(1 - I)$ decreases to zero and – since nobody is punished – there are no costs of punishment for society. When social costs are considered, the costs of harm implied by the first term must thus be balanced against the costs of punishment of the second and third term. Therefore, in the presence of costs of punishment, the social costs of harm must be weighted against the social costs of punishment, and therefore social costs are minimized for $\tilde{e}^* > \tilde{e}_{\text{neutral}}$. This result is based on Rizzolli and Saraceno (2013).

Identity Errors

Lando (2006) introduced a distinction between *mistakes of act* and *mistakes of identity*. *Mistakes of act* happen when a defendant is judged deviating when in fact he was complying. These are adjudicative errors we have been focusing on so far, for which the main concern of the adjudicative authority is whether there actually was any deviation at all. Note that, in case of mistakes of act, the two errors are independent: an increase in wrongful convictions does not imply any change in the number of wrongful acquittals. Then there are *mistakes of identity*, by which in the presence of deviations that can be easily observed, such as a murder or a robbery in the context of crime, the wrong person can be incriminated for an act that actually did happen. These are the cases where the occurrence of the deviation cannot be denied and the authority is concerned with who committed the crime. Note that in this case the two errors for a given crime are linked, as the conviction of an innocent person implies the acquittal of the person actually responsible for it.

Suppose that at time t_1 there exists an exogenous probability $\beta_{i,g}$ that an defendant is sanctioned for a deviation that has already happened at t_0 and which the subject is not responsible for (a mistake of identity). This exogenous probability can vary depending on the decision of the defendant at t_1: it seems reasonable to assume that abstaining from a crime at t_1 reduces the

probability of a mistake of identity, so that $\beta_i \leq \beta_g$. Thus the returns from conforming at t_1 are $E\pi_I = y_0 - (1 - I)s - \beta_i s$, while the returns from deviating are $E\pi_G = y_0 + b - (1 - G)s - \beta_g s$. The threshold level of b which implicitly defines the conforming population is

$$\tilde{b}_{\text{identity}} = (1 - (1 - I) - G)s - (\beta_i - \beta_g)s \quad (5)$$

Inspection of Eq. 5 and comparison with Eq. 1 highlight the role of mistakes of identity vis-à-vis deterrence implicitly defined by $\tilde{b}_{\text{identity}}$. The first part is equal to Eq. 1, while in the second part, if $\beta_i = \beta_g$ as Lando (2006) hypothesized, then identity errors occurred at t_0 have no impact on deterrence at t_1. However, if $\beta_i < \beta_g$, then identity errors actually have a positive impact on deterrence. The reason is intuitive: the decision to deviate in t_1 triggers a net increase in the probability of being wrongfully convicted because of a mistake of identity. Of course this result is based on the assumption that the probability of identity errors in t_1 is determined exogenously, and it is not a function of $\tilde{e}$. Furthermore, identity errors impose a necessarily constraint between the input of wrongful acquittals and the output of wrongful identity convictions; Garoupa and Rizzolli (2013) show that once this constraint is considered, mistakes of identity have a net negative impact on deterrence.

Errors and the Precaution of Harm

Another main area where the role of adjudicative errors has been explored is tort law (see Png 1986; Lando and Mungan 2017, among others). The standard model of tort law substitutes the dichotomous choice between complying and deviating with a continuous choice about the level of activity/care. We will discuss the main implications below. However, some novel conclusions can be drawn also from applying the dichotomous choice model. In this framework, the defendant chooses between *conforming* to the prescribed standard of care or *deviating* and not taking any precaution. Since taking precautions is costly, we can interpret b as the opportunity cost of conforming (by deviating, the defendant saves b). Furthermore, the sanction is equal to the harm inflicted since the goal of the tort system is compensation, and the decision to conform only reduces the expected harm: when conforming, harm h_i is produced with probability α_i, while when deviating, harm h_g is produced with probability α_g, where $h_g > h_i$ and $\alpha_g > \alpha_i$. Adjudicative errors can occur in the usual way, and therefore, a risk-neutral defendant's returns from conforming are $E\pi_I = y_0 - (1 - I)\alpha_i h_i$, while the returns from deviating are $E\pi_G = y_0 + b - (1 - G)\alpha_g h_g$. All defendants for which $E\pi_i \geq E\pi_g$ will conform and therefore the threshold level of b which implicitly defines the conforming population is

$$\tilde{b}_{\text{care}} = (1 - G) \cdot \alpha_g h_g - (1 - I) \cdot \alpha_i h_i \quad (6)$$

By comparing Eq. 6 with Eq. 1, one immediately realizes that type-I errors have a smaller impact on the incentive to comply than type-II errors as $\frac{\partial b}{\partial (1-I)} = \alpha_i h_i < \frac{\partial b}{\partial G} = \alpha_g h_g$; this is because complying causes a smaller expected harm. Also social welfare changes as now also complying defendant causes harm. To find out the optimal $\tilde{e}$, we compute $\frac{\partial SW}{\partial e} = 0$ and thus

$$\begin{aligned} &\partial Z(\tilde{b})\alpha_i h_i + \partial(1 - Z(\tilde{b}))\alpha_g h_g \\ &= (\alpha_i h_i i(\tilde{e}) - \alpha_g h_g g(\tilde{e}))(\alpha_i h_i - \alpha_g h_g) = 0 \end{aligned} \quad (7)$$

Rearranging Eq. 7, we have that $i(\tilde{e}) = \frac{\alpha_g h_g}{\alpha_i h_i} g(\tilde{e})$, and since $\frac{\alpha_g h_g}{\alpha_i h_i} > 1$, the equality can be satisfied only for $\tilde{e}^* < \tilde{e}_{\text{neutral}}$. We can thus conclude that when defendants face a dichotomous choice between complying and causing a smaller expected harm and deviating and causing a larger expected harm, type-I errors impact deterrence less than type-II errors, and therefore welfare is maximized for a level of evidentiary standard smaller than the neutral one.

Precautionary Activities and Chilling of Desirable Behavior

In both the crime and the tort contexts, the choice of deviating causes social harm at least in

expected terms. Compliance causes no harm in the crime context, while it produces a smaller social harm in the tort context. In many situations, however, defendant compliance can have both harmful consequences and benign ones. One may think of the case of competition policy, where the threat of antitrust sanctions may discourage efficient, pro-competitive behavior; another case may be medical malpractice, where worries about false positives may prevent cost-effective care. Kaplow's (2011) model envisages a population that can engage in a harmful act that produces a private benefit as well as a negative externality and another population that can only engage in a benign act that produces no externality. The two types of act are initially indistinguishable to the authority, but the adjudicative procedure gives rise to an evidence signal e that is higher for harmful acts than for benign ones. As before, the authority sanctions subjects whose acts produce an evidence signal higher than a certain cutoff value $\tilde{e}$. However, now the expected sanction raises both the costs of the harmful act and that of the benign one, thus chilling desirable behavior. Kaplow (2011) shows that the optimal $\tilde{e}$ that equates the (falling) marginal benefits of deterring harmful acts with the (rising) marginal costs of chilling benign acts is such that $\tilde{e}^* > \tilde{e}_{\text{neutral}}$. Intuitively it is generally optimal to raise the sanction and simultaneously raise $\tilde{e}$, holding deterrence constant. In fact the only consequence of this policy is a reduction in chilling costs.

A similar model is proposed by Mungan (2011) where subjects can choose between *inaction* (precautionary activity) and *action*, and this second choice can produce *no externality* (desirable behavior) or a *negative externality* (harmful activity). The authority cannot distinguish with certainty whether the activity is harmful or benign but must rely on an evidence signal e and balance the usual errors' trade-off. The expectation of sanctions wrongfully imposed on desirable behavior induces subjects at the margin to switch over to precautionary activities. Again, Mungan (2011) shows that the optimal evidence threshold is such that $\tilde{e}^* > \tilde{e}_{\text{neutral}}$.

Judicial Errors When the Choice of Care Is Continuous

In the model presented so far, the defendant's choice between complying and deviating is dichotomous. However, other situations like torts are best described by a continuous choice of care level x. In the prevailing model of tort, a legal standard $\dot{x}$ is set in order to determine liability by a potential injurer: the defendant avoids liability if his level of care is equal or above the standard one which is usually equated to the optimal level x^*. Craswell and Calfee (1986) introduce legal errors in this context by proposing a model where such legal standard is uncertain in the sense that defendants who choose a level of care x only know that there is a probability $F(x)$ (decreasing in x) that they will be sanctioned so that choosing higher levels of x decreases the probability to be punished. So if they choose $\underline{x} < x^*$, there is a $1 - F(\underline{x})$ probability of type-II error (the defendant is not made liable even if he took less than the efficient level of care), while if they choose $\bar{x} > x^*$, there is a $F(\bar{x})$ probability of type-I error (the defendant is made liable even if he took enough care). Assuming that also both the sanction s and the opportunity cost of care b are increasing functions of x, Craswell and Calfee (1986) show that with respect to the socially optimal level of x, the defendant's choice of x can be either undercomplying (the defendant chooses $\underline{x} < x^*$) or overcomplying $(\bar{x} > x^*)$. This is because, on one hand, there is always a positive chance $1 - F(x)$ of acquittal, and this increases the returns of taking lower levels of care. But on the other hand, the expected sanction depends on the probability $F(x)$, and this can be driven down by increasing the level of care. The relative impact of these two countervailing effects on the final level of x can be of either sign as it depends on various features of the legal environment and in particular on the amount of uncertainty. Craswell and Calfee (1986) show that under some plausible assumptions concerning the distribution of errors, defendants will usually take an excessive level of care. In other words, while the possibility of escaping liability when the defendant has not taken enough precaution (type-II error) has the usual adverse

effect on the incentives to take precaution, the possibility of being wrongfully held liable even when one has taken enough precautions (type-I error) induces the defendant to increase the level of precautions (under some plausible conditions).

Further Effects on Evidentiary Standards and on Type-I Errors

In addition to the literature survey above, many authors have also explained the high evidence threshold usually observed in legal trials using deterrence-based arguments that point at (i) biased evidence selection (Schrag and Scotchmer 1994), (ii) parties' evidence production expenditure (Yilankaya 2002), (iii) optimal exercise of care by parties (Demougin and Fluet 2006), (iv) marginal deterrence (Ognedal 2005), (v) repeated offenders (Chu et al. 2000), and (vi) emotional costs of indignation (Nicita and Rizzolli 2014).

Furthermore, without making reference to a specific utilitarian approach based on the deterrence rationale, a consistent number of papers simply postulate that wrongful convictions of innocents are morally repugnant and thus inherently worse than type-II errors. Arguments that justify this position are reviewed in Epps (2015). These arguments are mainly deontological and transcend the utilitarian framework (See also ▶ "Retributivism") although they can still be considered in our model by overweighting the impact of type-I errors on the social welfare function along the lines of Miceli (2009).

Empirical Relevance

The role of type-II errors has been greatly analyzed empirically and experimentally; there exists a vast literature testing Becker's deterrence hypothesis with real data on incarceration and on the death penalty (Chalfin and McCrary 2017). There is also a small stream of literature testing the deterrence hypothesis in the lab (see literature cited in Khadjavi 2015); (See also ▶ "Experimental Law and Economics").

Most of this literature, however, ignores type-I errors and their impact on deterrence and behavior. This asymmetry is easy to understand once one considers that type-II errors (crimes that go unpunished) are far more easy to be observed and measured than type-I errors (wrongful convictions can in fact be mistaken for correct convictions). Empirical research on type-I errors has only recently taken off either comparing agreement rates of judges and juries (Gould et al. 2014) or by using DNA testing introduced in the 1990s. Many innocent defendants used DNA testing to clear themselves after conviction whenever biological evidence from the crime scene had been retained. By adopting this strategy, Risinger (2007) estimated the type-I error rate in capital rape-murder cases to be between 3.3% and 5% in 1982–1989. Gross and O'Brien (2008), using post-1973 US data on death sentences (See also ▶ "Death Penalty"), estimated a type-I error frequency of wrongful death sentences to be at least 2.3%. Most of this literature is concerned with measuring the magnitude of type-I errors and less with the assessment of the impact of type-I errors on general deterrence (Gould et al. 2014). A small number of controlled lab experiments try to assess the impact of type-I errors on deterrence: Grechenig et al. (2010) first showed that both errors greatly undermine deterrence in a voluntary contribution mechanism (VCM) type of game. Rizzolli and Stanca (2012) disentangled the effects of each type of error and found that type-I errors are more detrimental to deterrence than type-II errors. Marchegiani et al. (2016) found the same effect within a principal-agent setting, while Markussen et al. (2016) using a VCM design found instead the opposite effect: that type-I errors are less detrimental than type-II errors.

References

Blackstone W (1769) Commentaries on the laws of England, vol 4. Clarendon Press, Oxford

Chalfin A, McCrary J (2017) Criminal deterrence: a review of the literature. J Econ Lit 55:5–48

Chu CC, Hu S-C, Huang T-Y (2000) Punishing repeat offenders more severely. Int Rev Law Econ 20:127–140

Craswell R, Calfee JE (1986) Deterrence and uncertain legal standards. J Law Econ Org 2:279–303
Dekay ML (1996) The difference between Blackstone-like error ratios and probabilistic standards of proof. Law Soc Inq 21:95–132
Demougin D, Fluet C (2006) Preponderance of evidence. Eur Econ Rev 50:963–976
Dhami S, al Nowaihi A (2013) An extension of the Becker proposition to non-expected utility theory. Math Soc Sci 65:10–20
Epps D (2015) The consequences of error in criminal justice. Harv Law Rev 128:1065
Garoupa N, Rizzolli M (2013) Wrongful convictions do lower deterrence. J Inst Theor Econ 168:224–231
Gould JB, Carrano J, Leo RA, Hail-Jares K (2014) Predicting erroneous convictions. Iowa Law Rev 99:471–2299
Grechenig K, Nicklisch A, Thöni C (2010) Punishment despite reasonable doubt – a public goods experiment with sanctions under uncertainty. J Empir Leg Stud 7:847–867
Gross SR, O'Brien B (2008) Frequency and predictors of false conviction: why we know so little, and new data on capital cases. J Empir Leg Stud 5:927–962
Kaplow L (2011) On the optimal burden of proof. J Polit Econ 119:1104–1140
Kaplow L (2012) Burden of proof. Yale Law J 121: 738–859
Khadjavi M (2015) On the interaction of deterrence and emotions. J Law Econ Organ 31:287–319
Lando H (2002) When is the preponderance of the evidence standard optimal? Geneva Pap Risk Insur Issue Pract 27:602–608
Lando H (2006) Does wrongful conviction lower deterrence? J Leg Stud 35:327–338
Lando H, Mungan MC (2017) The effect of type-1 error on deterrence. Int Rev Law Econ 53:1
Marchegiani L, Reggiani T, Rizzolli M (2016) Loss averse agents and lenient supervisors in performance appraisal. J Econ Behav Organ 131:183–197
Markussen T, Putterman L, Tyran J-R (2016) Judicial error and cooperation. Eur Econ Rev 89:372–388
Miceli TJ (2009) Criminal procedure. Edward Elgar Publishers, vol 3 of Criminal law and economics – encyclopedia of law & economics, Edward Elgar (ed), Chaltenham
Mungan M (2011) A utilitarian justification for heightened standards of proof in criminal trials. J Inst Theor Econ 167:352
Nicita A, Rizzolli M (2014) In Dubio Pro Reo. Behavioral explanations of pro-defendant bias in procedures. CESifo Econ Stud 60:554. ift016
Ognedal T (2005) Should the standard of proof be lowered to reduce crime? Int Rev Law Econ 25:45–61
Png IPL (1986) Optimal subsidies and damages in the presence of judicial error. Int Rev Law Econ 6:101–105
Polinsky A, Shavell S (1992) Enforcement costs and the optimal magnitude and probability of fines. J Law Econ 35:133–148
Risinger DM (2007) Innocents convicted: an empirically justified factual wrongful conviction rate. J Crim Law Criminol 97:761–806
Rizzolli M, Saraceno M (2013) Better that ten guilty persons escape: punishment costs explain the standard of evidence. Public Choice 155:395–411. https://doi.org/10.1007/s11127-011-9867-y
Rizzolli M, Stanca L (2012) Judicial errors and crime deterrence: theory and experimental evidence. J Law Econ 55:311–338
Schrag J, Scotchmer S (1994) Crime and prejudice: the use of character evidence in criminal trials. J Law Econ Org 10:319–342
Shavell S (1987) The optimal use of nonmonetary sanctions as a deterrent. Am Econ Rev 77:584–592
Yilankaya O (2002) A model of evidence production and optimal standard of proof and penalty in criminal trials. Can J Econ 35:385–409