E

Ecolables: Are they Environmental-Friendly?

Lisette Ibanez
LAMETA, CNRS, INRA, Montpellier SupAgro, Université de Montpellier, Montpellier, France

Abstract

This article provides a general overview of the technical, economical, regulatory and environmental aspects of ecolabeling. An ecolabel is a market-based policy instrument that can be either voluntarily adopted or mandated by law. Ecolabels are applied to services and products in order to inform consumers of their environmental-friendliness and to avoid market failures. In reality, however, ecolabels do not always succeed in achieving environmental improvements. The mis-use of environmental standards, the practice of strategic manipulations that create trade-distortions, the excessive use of claims, and behavioural biases are some of the factors that can prevent an ecolabel from reaching its initial objective to reduce or even eliminate environmental externalities.

History

The phenomenon of ecolabeling had its beginning in the 1970s, driven by an increasing consumer willingness to pay for sustainable consumption. Until that time, the management of environmental externalities relied mainly on the use of regulatory mechanisms that imposed requirements or restrictions on the operations of potentially polluting firms and entities.

The first official ecolabeling program, Blue Angel, was launched in Germany in 1977 and was followed by many other ecolabeling schemes in various countries and areas. Today, its development has permeated many sectors such as the food industry, cosmetics, home building, and even the automobile industry. After the 1992 UN Earth Summit conference in Rio, in which no worldwide agreement was reached for environmental protection, various initiatives (by governments, NGO's, private stakeholders, etc.) aimed to integrate environmental issues into manufacturing procedures. The idea was to influence consumption patterns in order to achieve sustainable development.

Since then, a burst of interest emerged for self-regulation and new environmental policy instruments, including ecolabeling, and an important change in manufacturing was observed in the United States: The percentage of new products claiming to be environmentally friendly rose from 1.1% in 1986 to 9.5% in 1999. Today, ecolabels are found everywhere: on a large variety of products and services, appearing on packaging, manufacturer's websites, or related to countries, or geographical areas. In total, 458 official ecolabeling schemes concerning 25 industry sectors have been reported in 195 countries

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

(Cf. http://www.ecolabelindex.com/consulted on October 2, 2014.). What's more, a great number of unofficial environmental claims are currently in use, as well.

Definition

Broadly speaking, ecolabels are claims related to environmental friendliness. They inform consumers about the environmental quality of products and services, enabling consumers to make better choices (Thogersen et al. 2010). It is widely understood that people rely heavily on information provided by labels in making their consumption decisions. Nevertheless, the word ecolabel remains a fuzzy and ill-defined term that can encompass a variety of different meanings. It includes a diversity of environmental claims ranging from third-party certification schemes to self-declaratory statements. Thus, at one extreme, ecolabels are issued by independent organizations (and can be voluntarily adopted by firms), as for example, Forest or Marine Stewardship Council Certifications (FSC and MSC). At the other extreme, ecolabels can be simple marketing claims or logos related to some aspect of environmental friendliness and are often vague, undefined, unverified, and/or unverifiable. Examples of these types of ecolabels include statements on detergent boxes indicating that the product contains no phosphates, or a label on a washing machine claiming that it is energy efficient.

In general, the conceptual design of ecolabeling schemes includes four stages (Grolleau et al. 2007). The first stage involves the selection of product categories or services to which the ecolabel will apply. The second stage defines the criteria that the product, services, or production process must respect in order to be allowed to use the ecolabel. Both of these stages, i.e., choice of product categories and ecolabel criteria, may be established by governments, environmental organizations, consumer groups, industries, or some combination of these parties. The third stage concerns the inspection of the products and services according to the defined criteria. This stage establishes the methods by which eligible products and services are to be tested and/or audited. The final stage develops the communication strategies or the way in which the environmental attributes will be signaled to the public.

Of course, the seriousness and accuracy of each of these stages vary highly according to types of environmental claims and labels and can therefore result in a wide range of compliance costs. In general, ecolabeling costs include the opportunity costs associated with efforts to conform to relevant standards (and thus testing and labeling) as well as the costs associated with efforts to create new markets. This means that firms must bear the costs of adjusting production processes in order to assure that products will earn the ecolabel, as well as the expenses of subscribing to and maintaining participation in these ecolabeling programs. Fees for these programs can be significant. Thus, under the assumption that firms are rational profit maximizers, their voluntary adoption of an ecolabel is conditional on the prospect of sufficient sales, higher prices, and due recognition of their environmental status by stakeholders and users.

Classification of Ecolabels

There are various ways of classifying ecolabels: they may be classified according to their legal status (mandatory or voluntary), the certifying entity (governmental, nongovernmental, or private rule-setting), the certification criteria (single or multiple criteria, product or process oriented, fixed or evolutionary mechanism), and even the geographical dimension of their production (regionally, nationally, or internationally defined).

In the interest of increasing the clarity and transparency of the ecolabeling process, the International Standard Organization has undertaken efforts to standardize ecolabeling principles. Specifically, it distinguishes between three types of environmental product labels, classified as ISO Type I, II, and III (Lathrop and Centner 1998). These three types of labels can be described as follows:

Type I (ISO 14024) labels are based on a set of criteria defined by private or public environmental

labeling programs and are issued and controlled by third-party certifiers. The awarding body may be either a governmental organization or a private noncommercial entity. Examples of such types of labels include the EU Ecolabel, the German Blue Angel, and the Program for the Endorsement of Forest Certification (PEFC). Type I labels are often referred to as ecolabels, despite the fact that they may coexist with other ecolabeling schemes (like the FSC program for example).

Type II labels are self-declared claims adopted by manufacturers, retailers, or anyone else likely to benefit from such claims. These claims can be factual (e.g., indicating the percentage of a product that is composed of recycled materials) or more unsubstantiated (e.g., simply "eco-friendly").

Type III labels consist of quantitative product information based on life cycle impacts. In general, these types of labels consist of reports that are presented in a form that facilitates comparison between products, often using a common set of parameters. However, relative product performance based on these parameters is not provided in the label itself, as there are no comparative statements made with respect to other products. An example of Type III claims is the "carbon footprint."

Despite their widespread use, relatively few regulations exist regarding the use of ecolabels on national levels. From a legal point of view, environmental claims are primarily seen as advertisements and less as market-based policy tools. As such, they are mainly regulated through antifraud and truth-in-advertising laws. However, the World Trade Organization (WTO) monitors the practice of ecolabeling in order to ensure that it doesn't lead to discrimination between trading partners or between domestically produced goods and services and imports. In the case of disputed environmental claims, the WTO relies on rules such as the General Agreement on Tariffs and Trade (GATT) or the Agreement on Technical Barriers to Trade (TBT). More specifically, the WTO issued a Code of Good Practice for Preparation, Adoption, and Application of Standards to TBT Agreement in order to facilitate ecolabeling and standard setting in a manner that does not conflict with agreed upon international trade frameworks. Often, it is very difficult to establish objective and scientifically defensible criteria that can justify or dispute claims of environmental superiority.

In cases where the voluntary involvement of an industry in an ecolabeling scheme is considered to be too low, governments may decide to oblige firms to undertake measures to reduce environmental impacts, adopt specific technology changes, or instate mandatory (national) labeling standards. For instance, in order to promote health, food, and environmental safety, governments may impose warning labels (e.g., for flammable materials, potentially dangerous household chemicals, alcoholic beverages, cigarettes). In these cases, producers who do not comply with the national standards cannot obtain access to markets within the country. For example, since 2010, EU Timber regulations prohibit the selling of illegally harvested timber and products derived from illegally harvested timber on the EU market. Trade barriers are not only enforced by laws, however, and can also result from consumer boycotts. Nevertheless, labeling requirements and practices must respect all the WTO regulations, including the provision of the TBT Agreement and the GATT. In other words, ecolabeling schemes should not discriminate either between trading countries and firms or between domestic and foreign products and services. There are several examples of debates regarding the legitimacy of ecolabels, due to the trade distortions arising from the enforcement of their environmental criteria. First, as stated by the OECD (2002) *"The experience of Colombian exporters of cut flowers to Germany provides an example of the impact which a powerful domestic NGO-driven voluntary eco-label can have on a developing country's trade prospects (OECD 2002)."* This conflict originated with German NGO's that campaigned against labor and environmental conditions in the flower export industry of developing countries. The Flower Label Program was created in 1996 and disputed by Colombian flower exporters who claimed that it created discrimination and unfair competition.

Another case in which pressure by local consumers and the media effectively excluded foreign

products from the domestic marketplace and created a conflict between international trade and domestic environmental law was the US/Mexico Tuna embargo dispute. In the 1990s, the US banned tuna imports that could not attest to being caught in a dolphin-safe way. The dolphin-safe label has legal status in the United States under the Dolphin Protection Consumer Information Act but was contested by Mexico as creating discrimination and illegal trade barriers (Dyck and Zingales 2002).

Principles of Ecolabeling

The rationale behind ecolabeling is to overcome the market failures that arise in marketing eco-friendly products and services. Markets for environmentally friendly products and services are dysfunctional because most of the promised environmental attributes are both public (nonrival and nonexclusive) and unobservable (credence attributes).

Neoclassical economic theory predicts that consumers will attempt to freeride by enjoying a public good (environmental quality) without incurring the costs associated with its provision. This means that if ecolabeled products are more expensive, there should be no market for them. However, in the last decades, consumers have increasingly considered not only their own well-being but also the environmental and social impacts of their purchasing decisions (Cohen and Viscusi 2012). Motivations for adopting socially appropriate behavior can be explained by preexisting personal norms or by emerging social norms. Nevertheless, even if it has largely been shown that consumers possess other-regarding preferences and are willing to pay for environmental attributes, another requisite for firms to invest in environmentally friendly production technologies and to offer less polluting products and services is the opportunity to provide credible information to users and consumers.

In order to determine the value that consumers attribute to an ecolabel, it is important to distinguish consumer understanding of the ecolabel, consumer confidence in the ecolabel, and their willingness to pay for environmental amenities (Delmas et al. 2013).

In the literature, it has been shown that simple ecoseal type labels are insufficiently convincing. Adding additional information about specific criteria that are defined and verified by a certification scheme can significantly increase willingness to pay (O'Brien and Teisl 2004). Ecolabels have also been found to perform better if the certifier is a familiar organization and/or endorsed by an independent third party (D'Souza et al. 2007), without ignoring that the success of the ecolabeling scheme is also determined by the individual characteristics of the purchaser (Teisl et al. 2008). Moreover, consumers are willing to pay higher premiums for ecolabels where the benefits of environmental improvement are also linked with private benefits. For instance, organic farmers strive to make their food products more appealing by stressing not only the reduced environmental impact of using less pesticides but also by emphasizing the accompanying reduced health risks.

Under the assumption that ecolabeling schemes have the ability to correct informational asymmetry within the producer-consumer relationship, firms are generally motivated to adopt ecolabels in order to differentiate their products. As pointed out by Porter (1991), (environmental) competitive advantages are obtained either by lower costs or by attribute differentiation. In many cases, new technologies with lower environmental impacts are not associated with lower costs; thus, a firm's primary interest in using ecolabels is founded in a search for strategic environmental competition. In developing a strategic advantage based on environmental quality attributes, firms can either focus on their internal organizational process or on their product line. Globally, differentiation will be based on creating a positive image, pursuing positive recognition by stakeholders, or the anticipation that more stringent environmental standards may be implemented in the future.

Voluntary label adoption as a differentiation strategy will affect competition, as it enables products to be perceived as imperfect substitutes (Bonroy and Constantatos 2015). In addition,

ecolabel adoption only partly corrects market failures relating to environmental degradation (Ibanez and Grolleau 2008). The incentives for firms to turn to new environmentally friendly standards and to adopt ecolabeling depend on several factors that can be classified into three general categories: factors relating to market structure and governance, willingness to internalize environmental externalities, and the ability of ecolabels to accurately communicate the certification criteria.

Firstly, there are factors related to the market structure and to the involvement, number, and type of policy makers. Firms' incentives to ecolabel are highly dependent on the costs of investing in green technologies, which concern both fixed (sunk) and relative unit costs, and on the ability of ecolabeling to release price competition. The inherent aim of the investment decisions made by firms is to consider how to best adapt their technology to standard requirements and do not necessarily aim at socially optimal outcomes in terms of environmental improvements (Amacher et al. 2004). Ecolabeling programs differ crucially from standard requirements in their ultimate goal. If an NGO initiates an ecolabeling program, it may likely set standards that maximize environmental improvements, whereas an industry-initiated ecolabeling program may instead be oriented by the pursuit of maximizing industry profits. Consequently, one would expect ENGO's to set more stringent standards than industry-sponsored ecolabeling schemes. It should be noted that more stringent environmental requirements do not necessarily lead to better global outcomes in terms of environmental degradation due to restricted use and consumption (Grolleau et al. 2009). Certifying entities may also compete to capture a firm's willing to adopt an ecolabel, strategically choosing their qualifying criteria in order to maintain their market share of participating firms. The coexistence of both NGO and industry ecolabels should theoretically favor industry profits, since environmental improvement is only feasible if the reduction in "high-environmental quality" labels doesn't outweigh the increase of overall ecolabel users (Fischer and Lyon 2014). Governmental involvement also plays a role in firms' willingness to adopt ecolabels and thus the overall effectiveness of ecolabels as an environmental policy strategy. In some cases, a market-based ecolabeling mechanism may be reinforced by complementary policy instruments in order to internalize externalities more efficiently (Nunes and Riyanto 2005). These complementary regulatory policies can either be mandatory (e.g., minimum environmental quality standards) or voluntary (e.g., subsidies for green investment).

A second determinant that explains the emergence of ecolabels concerns consumer willingness to pay for the reduced environmental impacts of products and services. The inclusion of social and environmental considerations in consumer decision-making processes has become more common in the last decades, and globally it has been shown that there exists a willingness to pay for social and environmental attributes. However, disagreements remain regarding the price premiums of these attributes (McCluskey and Loureiro 2003). This variation in estimated willingness to pay (WTP) observed in the literature can be attributed to the elicitation methodology used (based on either stated or revealed preferences) but can also, and is most often, structurally based. Indeed, ecolabeling may suit certain products better than others. As an illustration, if consumers value ecolabels for conspicuous reasons, it is important that the ecoseal and the associated product are seen by others. If warm-glow motivations are an important driver of behavior, then consumption of goods with symbolic logos (e.g., dolphin-safe fish, panda bear logo of WWF) may benefit from higher premiums. Thus, the design of an ecolabeling scheme and the associated communication strategy should consider the specific motivations underlying the demand for social and environmental attributes.

Lastly, the salience and efficiency of ecolabeling schemes rely on factors related to the ability of ecolabels to provide information on environmental attributes perfectly and symmetrically. Mason (2011) argues that incentives for firms to apply for ecolabels depend on the ability of third-party certification schemes to accurately distinguish high-environmental quality producers from low-environmental quality ones. Imperfect

label attribution arises either because monitoring is random or because the auditor is incapable of accurately identifying compliance with the required standards at a reasonable cost. This "noise" in testing procedures may even encourage less environmentally sensitive firms to apply for certification, which lowers consumer expectations of environmental quality in all similarly labeled products. Moreover, trust in an ecolabel is reinforced when the label is provided by independent and well-recognized certifying agencies (D'Souza et al. 2007) and if the entity has no financial interest in providing the certification (Harbaugh et al. 2011). In addition, the degree of consumer uncertainty regarding ecolabel claims rises along with increased market proliferation of these types of claims, as consumers will have more difficulty interpreting the absence or presence of an ecolabel, and consequently willingness to pay will decrease.

Efficiency of Ecolabels: Positive and Negative Side Effects

From a theoretical point of view, the introduction of ecolabels should be welfare enhancing as it provides information on the environmental attributes of products or services for which consumers are willing to pay, enabling simultaneous environmental improvements (Ibanez and Grolleau 2008). However, market-based policies and information disclosure, even if they are well known to minimize market distortions, may also result in unintended side effects. Indeed, the misuse of environmental standards, strategic manipulation that creates trade distortions, the excessive use of claims in general, and behavioral biases are some factors that might prevent an ecolabel from reaching its initial objective to reduce or even eliminate environmental externalities.

The motivations for introducing ecolabels and the process of defining and choosing standards are not necessarily the same among environmental organizations, governments, and industrial-oriented entities. While environmentalists focus on criteria that lead to the greatest environmental improvements, industries aim to set standards on product categories and/or criteria that involve the lowest costs and to take advantage of positive spillover effects such as an improved reputation among consumers and/or an increase in sales of other product categories (Dosi and Moretto 2001). Then, in some cases, the consequence of ecolabel introduction may in fact turn out to be globally harmful for the environment. In these cases, the environmental gains achieved are inferior to accompanying increases in product consumption and resulting pollution.

However, nonprofit labeling schemes do not always eliminate inefficiencies either. Potential examples include technological inertia in the form of reluctance to accept innovative technology developments, imposing excessively stringent standards or costly monitoring processes, and failing to gain sufficient recognition by consumers.

Moreover, the choice of labeling criteria can potentially be used for protectionist purposes. For example, industrialists may select environmental considerations that both meet environmentalists' concerns and disadvantage rivals. An intuitive example can be found in the environmental impacts of transportation. Domestic firms may overemphasize the role of certain transportation means (e.g., air transport vs. rail transport) in order to discredit foreign rivals. Differences in environmental issues among countries may also be employed in order to disadvantage foreign producers. Furthermore, for variety of reasons – e.g., political support, ideological protectionism – domestic governments may be more sensitive to arguments emanating from domestic producers and environmental activists, regardless of their scientific validity.

Harmonizing ecolabel standards on an international level seems to be a difficult task. Institutional, technological, and environmental heterogeneity between manufacturers is likely to favor strategies that seek to raise rivals' costs. In other cases, entities may focus on local or regional environmental priorities that lack international relevance and/or create trade barriers.

In general, developing countries are more vulnerable to the discriminatory impacts that can arise with ecolabeling schemes. First, it is more

difficult for these countries to bear the costs of the certification procedures required for the compliance, testing, and verification of the labeling criteria. Second, developing countries often lack information on requirements and do not have access to the necessary skills and technologies that would allow them to conform to domestic standards.

The proliferation of ecolabels also raises important questions. Theoretically, providing information on products and services should enable consumers to update their beliefs, thus having a positive impact on environmentally friendly consumption decisions. However, the proliferation of ecolabels today has, in some cases, led to an excess of information that users must process, contributing to increasing consumer confusion, indifference, and even skepticism towards ecolabeling schemes.

The overuse of environmental claims by firms may also be viewed as greenwashing by consumers. Consequently, there seems to be a need for more efficient auditing of the various stages of the ecolabeling process (Lyon and Maxwell 2011). But such regulatory intervention is costly, often imperfect, and not necessarily legally binding. Today, no binding legal framework at the international level exists in order to ensure the reliability of green claims, i.e., in order to verify whether the relevant labeling criteria are respected as defined by ISO 14021. To fill this gap and to call for a legal framework, private initiatives have denounced firms that use "misleading" and/or inaccurate advertising regarding their actions in favor of the environment. In 2008, several environmental NGO's decided to attribute an award, called the Pinocchio prize, to firms who promoted themselves to be greener than they actually are (http://www.prix-pinocchio.org/).

In order for ecolabels to be effective, consumers must receive, process, and believe the information transmitted by labels. Therefore, the type and the content of information transmitted by ecolabels are of crucial importance. The aim of an efficient label design is to provide consumers with information that increases understanding of environmental targets and improvements in a credible manner and may simultaneously boost pro-environmentally friendly preferences themselves (Delmas et al. (2013)).

Truthful ecolabels do not necessarily achieve environmental policy goals. Because consumers tend to focus only on the emphasized attribute, they often underestimate the overall environmental harm resulting from their consumption and use of the product. For instance, by buying fluorescent light bulbs, consumers may focus on the energy saving properties yet underestimate the environmental damages that result from the emission of mercury vapor occurring when these lights are disposed. A great many other behavioral biases may also distort the efficiency of ecolabeling schemes (Grolleau et al. 2015). For example, people in general overestimate their own virtuous behavior (optimism bias) and underestimate the degree to which they are implicated in contributing to negative externalities (attribution error). If these cognitive biases apply to ecolabeled products, people may allow themselves to increase consumption of environmentally friendly products thinking that their consumption is less harmful, while in reality their increased consumption may in fact contribute to a net deterioration of the environment (Bougherara et al. 2005).

An important determinant in consumer behavior is other-regarding and social preferences. And one can imagine the progressive shift of markets towards more sustainable production and ecolabeling may have the potential to increase consumer preferences for environmental-friendly attributes in the long run. In other words, exposure to an increasing and diverse range of ecolabels can trigger new social norms pertaining to acceptable environmental behavior.

However, even if exposure to green products may activate norms of social responsibility and ethical conduct, it may also give rise to a rebound effect (Cohen and Viscusi 2012). Recent research has shown that green consumption or pro-environmental conduct can modify people's overall sense of morality and affect subsequent behavior either within or outside of the concerned domain (Clot et al. 2014). This moral compensation, also called the "licensing effect," shows that doing something virtuous seems to give individuals license to later engage in more immoral

actions. The consequences of this phenomenon can be diverse and have negative impacts on the global environment through the reduction of pro-environmental behavior in subsequent decisions or in other domains.

Conclusion

In theory, ecolabeling is an efficient way of regulating environmental externalities, under the condition that the additional costs don't exceed consumer willingness to pay for environmental quality. When this is the case, voluntary approaches for environmental regulation are good alternatives to command and control policies. However, in real life, ecolabeling does not always achieve its initial goal of environmental improvements due to a variety of side effects. Intervention or regulation could then provide a way to correct these inefficiencies (Horne 2010).

But what should be the role of governments? Is new legislation the only way to regulate inefficiencies, or can efficient voluntary mechanisms be implemented to effectively encourage optimal behavior? It is important to be able to measure the global effectiveness of ecolabeling programs through assessments on an international level. The importance of environmental improvements in relation to trade distortion raises ethical and equity issues and calls for better insights on the economic implications of the interaction between ecolabeling schemes and (already existing) command and control policies.

Regarding the extent to which ecolabeling should be regulated, a distinction should be made between interventions in the certification process and scheme, and interventions regarding the use of ecolabels as a communication tool. A first task is thus determining which products and services should be targeted to use ecolabels. One might wonder whether consumers should systematically be encouraged to purchase ecolabeled products and services and firms pushed to adopt ecolabels and more sustainable production technologies. Complementary policies on both the national and international level should also target ecolabels. Nevertheless, the simple provision of accurate product information is often not enough to induce behavior change, and thus social and behavioral research could help policymakers to better design ecolabels in order to improve their overall effectiveness (Grolleau et al. 2015). One promising new strategy emphasizes the importance of educating the youngest generations in environmental issues and the relevance of ecolabels to these issues.

References

Amacher G, Koskela E, Ollikainen M (2004) Environmental quality competition and eco-labeling. J Environ Econ Manag 47:284–306

Bonroy O, Constantatos C (2015) On the economics of labels: how their introduction affects the functioning of markets and the welfare of all participants. Am J Agric Econ, forthcoming 97(1):239–259

Bougherara D, Grolleau G, Thiébaut L (2005) Can labelling policies do more harm than good? An analysis applied to environmental labelling schemes. Eur J Law Econ 19:5–16

Clot S, Grolleau G, Ibanez L (2014) Do good deeds make bad people? Eur J Law Econ, forthcoming https://doi.org/10.1007/510657-014-9441-4

Cohen M, Viscusi WK (2012) The role of information disclosure in climate mitigation policy. Climate Change Econ 3(4):1–21

D'Souza C, Taghian M, Lamb P, Peretiatko R (2007) Green decisions: demographics and consumer understanding of environmental labels. Int J Consum Stud 31:371–376

Delmas M, Naim-Birch N, Balzarova M (2013) Choosing the right eco-label for your product. MIT Sloan Manag Rev 54(4):10–12

Dosi C, Moretto M (2001) Is ecolabelling a reliable environmental policy measure? Environ Resour Econ 18(1):113–127

Dyck A, Zingales L (2002) The corporate governance role of the media. Chap. 7. In: The right to tell. The role of mass media in economic development. World Bank Institute, Washington, DC, pp 107–140

Fischer C, Lyon T (2014) Competing environmental labels. J Econ Manag Strateg (forthcoming) 23(3):692–716

Grolleau G, Ibanez L, Mzoughi N (2007) Industrialists hand in hand with environmentalists: how ecolabeling schemes can help firms to raise rivals' costs. Eur J Law Econ 24(3):215–236

Grolleau G, Ibanez L, Mzoughi N (2009) Too much of a good thing? Why altruism can harm the environment. Ecol Econ 68(7):2145–2149

Grolleau G, Ibanez L, Mzoughi N, Teisl M (2015) Helping eco-labels to fullfil their promises, Climate policy https://doi.org/10.1080/14693062.2015.1033675

Harbaugh R, Maxwell J, Roussillon B (2011) Label confusion: the Groucho effect of uncertain standards. Manag Sci 57(9):1512–1527
Horne R (2010) Limits to labels: the role of eco-labels in the assessment of product sustainability and routes to sustainable consumption. Int J Consum Stud 33:175–182
Ibanez L, Grolleau G (2008) Can ecolabeling schemes preserve the environment? Environ Resour Econ 40(2):233–249
Lathrop K, Centner T (1998) Eco-labeling and ISO 14000: an analysis of US regulatory systems and issues concerning adoption of Type II standards. Environ Manage 22(2):163–172
Lyon T, Maxwell J (2011) Greenwash: corporate environmental disclosure under threat of audit. J Econ Manag Strateg 20(1):3–41
Mason C (2011) Eco-labeling and market equilibria with noisy certification tests. Environ Resour Econ 48:537–560
McCluskey J, Loureiro M (2003) Consumer preferences and willingness to pay for food labeling: a discussion of empirical studies. J Food Distrubution Res 34(3): 95–102
Nunes P, Riyanto Y (2005) Information as a regulatory instrument to price biodiversity benefits: certification and ecolabeling policy practices. Biodivers Conserv 14(8):2009–2027
O'Brien K, Teisl M (2004) Eco-information and its effect on consumer values for environmentally certified forest products. J For Econ 10:75–96
Porter M (1991) Towards a dynamic theory of strategy. Strateg Manag J 12:95–117
Teisl M, Rubin J, Noblet C (2008) Non-dirty dancing? Interactions between ecolabels and consumers. J Econ Psychol 29:140–159
Thogersen J, Haugaard P, Olesen A (2010) Understanding consumer responses to ecolabels. Eur J Mark 44(11/12):1787–1810

Economic Analysis of Brazilian Labor Law

Sergio Pinheiro Firpo and Luciana Yeung
Insper Institute of Education and Research, São Paulo, Brazil

Definition

Being a civil law country, Brazil heavily relies on codified rules to regulate labor relations. For this purpose, back in the 1940s, a comprehensive set of laws was consolidated in the *Consolidação das Leis Trabalhista* (CLT), which remains the main source of labor regulation in the country, even after some minor reforms in the end of the year 2017 (to be discussed ahead). Besides the CLT, the Federal Constitution itself has several clauses governing labor and employment relations. Finally, more and more judicial precedents have been taken and followed (although not in a mandatory manner) as sources of law.

Introduction

Labor law in Brazil is based on a paternalistic, overly protective, and dictatorial view of employees. The first and perhaps most important principle is that of the vulnerability of employees: these are considered almost legally incapacitated, with no knowledge whatsoever of what their desires are. This is the first and sharp contrast with the economic concept of rational agents: people do know what their preferences are and what costs they must incur in order to maximize their benefits.

Another assumption by the Brazilian labor law is that of adversarial relations between employers and employees. On one hand, employees are considered to be irrational, not knowing what their preferences are, and totally lacking bargaining power; on the other hand, employers are considered to be pure profit maximizers, willing to adopt inhumane conditions for his/her employees with the sole objective to reduce production costs. Because of that, legislators, judges, and law enforcers will stay vigilant to regulate this relationship. Under such circumstances, the scenario predicted by Coase, of cooperative bargaining, leading to efficient allocation of resources, becomes even more a piece of fiction. Labor relations, by definition and assumption, are considered to be of very high transaction costs. Then, as Coase's theorem clearly predicts, the law will directly impact on the efficient allocation of resources. The question is as follows: Has this impact been mostly positive or mostly negative in Brazil? The answer is not difficult to find. We provide in the second half of this

entry a brief review of the empirical literature showing the negative impacts on efficiency of these regulations.

Main Characteristics of Labor Law and Labor Justice in Brazil

Conflicts are pervasive in judicial relations in Brazil. The judiciary power is comprised of five big branches: Federal justice (for all matters related to the Federal Constitution), common state justice (for most civil cases, such as commercial, tort, family, consumer rights, etc.), military justice (for trials involving the military), electoral justice (running the electoral system), and labor justice. Within it, a three-tier structure exists, comprising local first-instance courts, second-degree regional courts, and a supreme labor court. In cases where labor conflicts involve constitutional matters, cases and appeals may be directed to Federal courts, even the highest Federal Supreme Court (STF).

Given such a structure, it is not a surprise that this is a highly conflictive branch of law. Official statistics published by the National Council of Justice (CNJ) – the "Justice in Numbers" (*Justiça em Números*) reports – testify this: in 2016, there were 1721 new cases in labor courts for every 100,000 people, or a total of 4.3 million new cases. With regard to pending cases, labor courts comprised 5.4 million, placed top 3 (6.8% of total), behind only to state courts and Federal courts. As a specialized branch of the judiciary, labor courts have far more cases than those of the electoral justice, which in 2016 received only 972,000 new cases.

Furthermore, "Justice in Numbers" also shows that the most frequent theme in all Brazilian courts, the top 1 subject of judicial conflicts, is "labor dismissal/dismissal fees," comprising 11.5% of all cases brought to courts in Brazil. In 2016, they represented more than 5.8 million cases. Top 2 conflictive subject was "breach of contracts" (not including labor contracts), with far less, 1.9 million cases.

One can affirm, with certainty, that labor laws are not bringing harmony in employers-employee's relations, or in Coasean terms, the law is not reducing transaction costs.

Historical Background: The Creation of Labor Laws and Labor Justice in Brazil

Labor law and labor justice were born under a dictatorial regime in Brazil. The first compilation of labor laws, the *Consolidação das Leis Trabalhistas* (CLT), was made in 1943, 2 years after the birth of labor courts in the country. Both are creations of populist dictator, Getúlio Vargas. In his turn, Vargas came to power with the 1930 revolution, which put an end in the "old republic" in Brazil, which had been dominated by the rural coffee aristocracy. Vargas' regime represented the beginning of the political power exerted by urban industrialists and bourgeoisie. His government was one of the extreme instabilities, pointed by several attempts of coup and a few effective ones, ending in a dramatic way, with the suicide of Vargas in 1954. He took power twice, in a total of 18.5 years (1930–1945 and, then, 1951–1954). It was during his first government, under a dictatorship, that labor law was created. It is usually considered to be a populist dictatorship, commonly observed in several Latin American countries at that historic time.

Those were years of intense social transformations in Brazil: industrialism, urbanization, and influx of foreign immigrants. In order to maintain social order, Vargas managed to control the various classes that were emerging, especially urban groups. Being a founding member of the "Brazilian Labor Party" (*Partido Trabalhista Brasileiro*, PTB), he had strong commitment with the labor class. Another important historical event from those years was the emergence of communist ideas, mainly brought by the Italian immigrants. In response to all these movements, the creation of the labor justice and of labor laws was a means to pacify and monitor potential labor movements adversarial to the government. The paternalistic approach adopted by this legal structure was a manner to compensate the

intense monitoring by the government that the new system established.

According to Lopez (1991), the unification of all previous labor laws under the CLT represented a "combination of state paternalism and fascism, which was the essence [of Vargas' dictatorship]" (free translation). Some characteristics of the Brazilian labor structure reveal the accuracy of this analysis. Vargas created a "fake" unionism, which included monopoly of representation, i.e., only one union per professional category in a particular geographical region. Employees must pay compulsory union fees (deducted from their paychecks once a year). In turn, unionism is funded by public resources. Unsurprisingly, unions were mostly subordinated to the Federal government. By its turn, labor laws, greatly represented by the CLT, impose high financial costs and responsibility burden on employers. Legal rules are generally applied under the rule of strict liability (Yeung 2006). Labor law is based on the principle of vulnerability of the employee, under all circumstances. This means labor conflicts will never be solved as contractual conflicts by Brazilian courts. There is a long list of labor benefits that employers are mandated to provide. This, in turn, increases labor expenses up to 200% of what employees effectively receive as paycheck (Pastore 2007).

Paternalism toward employees is also reflected in procedural rules. Until a very recent reform in labor laws, employees had no costs at all to access labor courts: all that was required was a self-declaration of financial incapacity. This certification was, then, later submitted for approval by the judge. Decisions were made based on the judge's subjective opinion; there was no requirement for financial statements. This procedural rule was granted by a combination of principles present in Brazilian labor laws, a rule issued in 1983 (Law Number 7115) and several labor jurisprudence. Furthermore, although provided by procedural law, "bath faith litigation," or litigation without reasonable motif, is rarely punished in labor courts. With all this in place, it is not difficult to understand why there is so much conflict in Brazilian Labor Justice.

Distortionary Effects of Labor Law in the Brazilian Labor Market

There are three important features of labor law in Brazil that, in spite of being the source of several distortions in the labor market, they were not affected by the 2017 reform, which we shall discuss in the next section. These three aspects are *firing costs*, *labor taxes*, and the existence of the Brazilian Severance Indemnity Fund, known as *FGTS* (Fundo de Garantia do Tempo de Serviço). These three components have had important effects on the labor market over the past decades affecting employment levels, unemployment and informality rates, and labor productivity.

Firing (or dismissal) costs are the set of rules in labor law that cause the firm-employee separation, whose initiative is by the employer, to involve some pecuniary penalty to the firm, both by means of indemnification to the dismissed worker and by means of fines paid to the government, in addition to prior notice to the worker. Firing costs in Brazil are relatively high, as they involve a fine of 40–50% of a monthly salary for each year in the firm, plus 1 month as prior notice to the worker, who may not need to fully work during that period. Higher dismissal costs induce lower job variability and a lower level of formal employment because firms tend to hire less, since any reversals in product demand could lead to costly layoffs in the future. The effects on informality are the opposite, and higher dismissal costs induce higher levels of informality. Finally, because (i) dismissal costs have a reallocation effect from formal to the informal sector and (ii) the latter has an average productivity below that of the formal sector, the higher the firing costs, the lower the labor productivity. In addition to that effect, there is the productivity-reducing effect in the formal sector, because firms in that sector will typically operate below the optimal labor demand curve.

Labor or payroll taxes are those that are paid when hiring an employee. There are many taxes and charges on payroll in Brazil, and they may vary between 20% and 50% of overall labor costs. These taxes are, in general, proportional to the time of service and to the salary paid. The effect of levying labor taxes is well known and amounts to the

reduction of employment. These taxes shift the labor supply curve inward and upward, and therefore their effects on unemployment and informality are inversely proportional to what they have on employment. Finally, labor taxes reduce the marginal productivity of labor for two reasons. First, after an increase in labor taxes, workers who will remain employed are those in the points along the supply curve that are associated with lower reservation wages. Because reservation wages reflect productivity, average labor productivity will decrease after an increase in labor taxes. The second reason is related to the increase in informal sector induced by these taxes, which tends to operate under lesser conditions of productivity.

The last important feature of the Brazilian labor law that has not been affected by the recent reforms is FGTS, the Brazilian Severance Indemnity Fund, which was created in 1966. From the 1940s until 1966, Brazilian labor law stipulated that in the case of unjustified dismissals, firms should pay as compensation to the worker one monthly salary for each year worked at the firm. In addition, the law guaranteed job stability for workers who had an employment spell longer than 10 years with the same firm. FGTS was introduced with the aim to introduce flexibility in the job relations. Since 1966, workers have not had job stability. But they have kept access to the one monthly salary for each year worked at the firm. That amount is deposited by the firm into an individual FGTS account. Workers also are eligible to a 40% fine over the balance in the case of unjustified dismissals that is paid by the firm at the time of dismissal. Thus, FGTS is not a simple labor payroll tax: it is a mandatory savings account paid by the employer plus a fine that is appropriated by the worker. That account can only be accessed if the worker is fired. Voluntary dismissals do not allow withdraws from the FGTS account. Therefore, FGTS induces the opportunistic behavior of workers who always benefit from being fired: they may have granted access to their FGTS account and to the fine. Because the interest rates accruing FGTS accounts are lower than inflation, it is rational for workers to try to have access to their individual accounts. The overall effects of FGTS are excessive labor turnover, which reduces labor productivity. Since it increases labor costs for the firms, it also has all those other effects already described for labor taxes: lower employment, higher unemployment, and informality rates.

Labor Law and Labor Courts After the 2017 Reform: An Economic Analysis

In November 2017, a so-called labor law reform took place in Brazil, 74 years after the creation of the CLT, which remained basically untouched during all this time. In fact, it was approved by the National Congress in July, under Law N. 13,467, in a very heated process, severely opposed by trade unions and labor movements. The law in its integrity has close to 900 articles. For the purposes of this paper, we highlight three very important ones. We then provide a brief economic analysis of these topics (Yeung 2017).

1. (Articles 611-A and 611-B) Individual and/or collective agreements gained much importance as compared to legally issued laws over the discussion of a wide variety of themes. For a civil law country, highly dependent on legislative rules, this was a significant change in the process of creation of formal rules for labor relations.

Economic analysis: This is one of the most controversial points of the 2017 labor reform. However, it is very much in line with an economic analysis of labor, especially under the concept of rational agents and of the Coase theorem: employees know what is good for them and, when bargaining conditions are guaranteed and transaction costs kept in low levels, the result of the cooperative bargaining between employees and employers will be efficient. This is especially true if negotiation occurs with the support of labor committees or labor representatives at the workplace (as predicted by Article 510-A). Furthermore, although Article 611-A predicts 15 matters over which cooperative agreements and private bargaining can stipulate, Article 611-B lists 30 matters (i.e., double the previous number) which there must be *no* private negotiation, over

which only the law may regulate. These are mostly laws related to situations in which negative externalities are created by the employer to the employees and to the society.

2. (Articles 545, 578, and 579) Employees are no more legally required to make contributions to trade unions: the rule that mandated each employee to make an annual contribution equivalent to the wage of 1 day of labor, stipulated by Getúlio Vargas, is finally banned. Now, in order to get the contribution, unions must have written consent by the employee (but the rule of monopolist trade unions was unchanged).

Economic analysis: The former rule mandating union contribution by each formal employee was a clear violation of the freedom of association, especially in the case of Brazil, where there exist monopolist trade unions. That law could only have been created under a dictatorial political regime. Economic analysis would go beyond in the normative recommendations and demand for the extinction of union monopoly.

3. (Art. 791-A) Sums of surrender are owed by the employee, in cases of judicial defeat. Curiously, this rule has normally been applied to all other civil cases, even those with legally vulnerable parties, such as consumers against enterprises. The only exception happened in labor courts, and now, with the 2017 reform, this exceptionality has also gone away.

Economic analysis: If employees have "nothing to lose" when accessing labor courts, certainly they will access. This is one main reason of the statistics with evidence of over litigation in the labor justice, as seen before. Another reason might be a potential bias pro-employee and anti-employer by labor courts. The combination of these two factors leads to the very adversarial relationships between the two parties. With the end of the gratuity to access labor courts, it is expected that there will also be a reduction in the numbers of labor litigation.

Few months after its real implementation, some anecdotal evidence show the 2017 labor reform has already produced concrete impacts on courts, by significantly diminishing the number of labor cases filed. We believe it is still early to significantly draw such conclusion. Hopefully, future academic research will address whether this small step was able to significantly improve the environment of labor relations in the country.

Cross-References

► Economic Analysis of Labor Law

References

Lopez LR (1991) História do Brasil Contemporâneo, 6th edn. Mercado Aberto, Porto Alegre

Pastore J (2007) Trabalhar custa caro. Editora LTr, São Paulo

Yeung L (2006) The need of modernizing Brazilian labor institutions. In: 14th World Congress – Industrial Relations Research Association, 2006, Lima, Peru. Social Actors, Work Organization and New Technologies in the 21st Century. Lima, Peru: Universidad de Lima – Fondo Editorial, vol 1, pp 117–133

Yeung L (2017) Análise Econômica do Direito do Trabalho e da Reforma Trabalhista (Lei N° 13.467/2017). REI-Revista de Estudos Institucionais 3(2):891–921

Economic Analysis of Judicial Decisions

► Empirical Analysis of Judicial Decisions

Economic Analysis of Labor Law

Luciana Yeung
Insper Institute of Education and Research, São Paulo, Brazil

Abstract

The literature on labor relations has been very much divided, mainly between the classical economic approach and the legal/sociological

one. An economic analysis of labor law would benefit by abridging both perspectives and complementing with more. This entry provides brief comments of why whether a purely economic approach on labor or a purely legal approach based on capital vs. labor dichotomy is not sufficient to properly address the subject. Several topics related to labor regulation are discussed, and empirical references on those issues are provided.

Economic Analysis of Labor Law and Labor Relations

Since the beginning of humankind, labor was exercised for people's needs. The study of human labor, its relation with the environment and its impacts in human relations within groups, is possibly one of the oldest social studies. Since the Greek philosophers and the Medieval religious scholars, up to the modern and contemporary economists, social scientists, psychologists, industrial engineers – just to name a few – several brilliant minds have engaged their time and research to understand the phenomenon of human, either in a positive or a normative manner. Many of the classical titles in economics and sociology were inquiries on topics related to the nature and consequences of human labor, such as Karl Marx's *The Capital*, Max Weber's *The Protestant Ethic and the Spirit of Capitalism*, and, in some sense, Adam Smith's *An Inquiry into the Nature and Causes of the Wealth of Nations.*

However, in recent times, the study of labor has been absolutely divided. On one side, economics – and specifically labor economics – has been mainly concerned about players' benefits. The economic approach is concerned with the analysis of workers' jobs, income and benefits (short term and long term), pension, etc. For the employers' side, economic models have studied impacts on productivity, profits, flexibility, etc. This perspective sees workers and employers as if the only things these actors pursue in the labor arena are economic and material benefits and that the main relationship here is of sellers and buyers. This is the basic framework of labor economics: a market in which demand (workers) and supply (employers) interact.

On the other side, there is a wide range of scholars analyzing labor relations (until very recently also known as industrial relations) in a very different perspective. This group includes labor sociologists, labor lawyers, and labor historians – among others, who do not view employers and workers as merely in a seller vs. buyer relationship; instead, these researchers see a naturally antagonistic relationship at the labor *arena*. The model of capital vs. labor is the axis of this analysis, and all outcomes derive from it. Especially in the case of labor law, it aims at deriving rules that would smoothen this conflict (whenever possible) or that would balance the opposing forces, normally by protecting the weakest part, i.e., workers. These scholars consider the *locus* of labor as a battlefield, in which conflictive relations pervade, as if confronting with each other was the only thing that matters to these actors in their daily encounters at the workplace.

Needless to say, both approaches are limited and insufficient.

An economic analysis of labor law should take both sides of the view and complement with more. It incorporates features of the demand-supply model of labor and especially recognizes that workers and employers equally face several kinds of *incentives* and *constraints*, which affect their decision-making. However, it also gives significant importance to the formulation, application, and enforcement of rules, whether contractual or regulative ones, in the labor arena. It recognizes that the relationship between employers and workers is not a mere one of seller vs. buyer, or supplier vs. enterprise, and that labor is not a simple commodity. Human relations and power relations matter much here.

Although economic analysis of law a priori adopts the main economic framework, which considers labor *locus* a market, it recognizes that this is a special one, in which failures are the norm: asymmetric information (either from the employer to the worker or from the worker to the employer), externalities, uneven bargaining and political powers, monopoly, and monopsony,

among others. Externalities are sources of high transaction costs, and as the normative approach of the Coase Theorem tells us, under these circumstances, legal rules have an important role in the determination of the levels of efficiency (Coase 1960; Cooter and Ulen 2007). In other words, in this market, institutions matter and matter a lot.

Special Rules for a Special Market

Market failures exist in all markets. Yet, in the context of labor, they are essential features, not exceptions.

Asymmetric Information

Whenever labor is done for others – i.e., not for one's own subsistence – it entails a contractual relation, either a formal or an informal one. It is a contract because it is a *promise* made by someone to deliver something in exchange of another thing, and this is an enforceable promise. Workers promise to work under such and such conditions, to deliver certain tasks and/or to obtain certain results, in exchange of salary and a certain set of benefits. Employers, in their turn, promise to provide certain working conditions in exchange of workers' labor, which is used to produce goods and services, which, in turn, generate receipts and profits. The presence or absence of a written paper does not alter the fact that this relationship has a contractual nature, because both formal and informal contracts have their valid mechanisms of enforcement; it does not mean that written contracts are less effective in the guarantee of productive outcomes. The comparison between formal and informal types of contract is not the main issue of this brief entry, and, for our purposes, one should only bear in mind that labor entails, indeed, a contractual relation.

Besides that, most of the transactions between employers and employees are non-instantaneous, long-term interactions. The longer the relationship, the stronger are the impacts of uncertainties. For instances, neither the employer nor the employee knows, at the time of the recruitment, whether the economy will be booming or will be slowing down in the following years (which may impact in the employer's ability to pay higher salaries); both parties do not even know how long will the company survive in the market.

Finally, human interactions are characterized by imperfect and unbalanced information: one is never able to access the whole set of information related to the other party, due to limited cognitive capability and due to high transaction costs. This is true even for noneconomic human relationships such as marriage, friendship, etc. In the case of employers, even if they try hard to access candidates' true ability (even by means of sophisticated processes of screening), they will not be able to fully acknowledge the candidate's adequacy for the position. Employers will also not be able to access the candidate's real interests in joining their firm ("Does this candidate plan to stay here long? Or will she/he leave the company as soon as another opportunity appears in the rival company?"). On the employee's side, she/he, at the time of the recruitment, is not able to fully access the challenges of the new job (even if one seeks information with current employees); she/he might not be able to understand the easiness or difficulty to interact with the new boss. Not even is the candidate able to access vital information about the company's finances and organizational challenges.

Externalities and Interdependence of Utility Functions

Employers depend on workers to reach some of their desired outcomes; workers, on their side, depend on employers to reach their economic benefits. This is what economist calls an *interdependence* of their utility functions. Whenever it happens, negative externalities might be created: one's decisions or actions might inflict undesired costs to the other party.

However, positive externalities might also be created. In the labor context, whenever an employer decides to train or educate his/her workers, he/she creates benefits not only for the company, for the trained workers, but potentially to the whole society, since a highly skilled labor force generates technology and knowledge that may benefit the whole society, directly or

indirectly. This is true both for general and for firm-specific skills. The same happens when employers invest in workers' health and safety conditions and so on. Thus, giving incentives for employers to invest in his/her employees' well-being also generates social efficiency. Yet, one knows from economic models that externalities – both negative and positive – drive the economy away from the point of efficiency: in the presence of *negative* externalities, private parties tend to provide *excessive* amounts of products and services; on the other hand, *positive* externalities lead to *insufficient* amounts. In both cases, regulation is necessary to guarantee efficiency: for negative externalities, taxes and quotas must be stipulated; for positive externalities, public policies must provide subsidies or other incentives to stimulate their production.

Uneven Bargaining and Political Powers

Economists are unable to deal with the concept of power relations. This is a phenomenon that cannot be explained by economic rules nor economic models. Yet, they significantly affect bargaining outcomes, even in the context of Coasean bargaining, with impacts on efficiency (e.g., Galiani et al. 2014; Barnes 2004, among others). It is also unrealistically naïve to consider this market as having "normal" supply and demand forces, with equal bargaining power. For most systems, employer and employees – especially if these ones are treated individually – do *not* have equal stands in the negotiation of working conditions. Thus, to assume that this is an ordinary contract relationship, in which clauses are outcomes of cooperative and voluntary agreement, does not truly mirror reality. Sociological, cultural, and political variables may explain why workers' bargaining powers are, on average, stronger in one system than they are in others, but the common-law tradition which views labor regulation as another ordinary type of contract regulation might not be well suited for other countries; probably, there are places in which power imbalance between workers and employers is larger. Then, other types of labor rules must be considered, so to take the "power effect" into consideration.

Monopoly-Monopsony

The historical answer to uneven bargaining and political powers in labor relations was labor organizations, mostly (but not exclusively) trade unions. Yet, this creates a problem of another sort: unions operate, most of the time, as monopolies in the labor market. As one knows, monopolies create deadweight loss: due to the existence of unions, wages tend to be higher, and as a response, employers are willing to hire less labor. In addition to that, because membership is never mandatory, the presence of unions creates an even worse outcome, dual labor markets, in which unionized workers have higher benefits but at the cost of nonunionized ones. In poor countries, this is materialized, in its extreme, into the case of formal vs. informal labor markets, with a large portion of the labor force belonging to the second one. Needless to say, working and living conditions here are strikingly adverse. We will discuss some empirical findings about trade unions in a specific section ahead.

It is also not unusual to find cases of monopsony in labor markets: one or a few firms are the sole employers of a certain type of workers in a region. Under these circumstances, employers have abnormal market power to unilaterally set wages, working conditions, etc., to which workers have little chances to react. A monopsony may be as damaging for social welfare as monopolies.

Litwinski (2001) discusses both problems in the labor market and, within the American context, defends the application of antitrust law to unions (as it has happened in the USA since the beginning of the twentieth century, in the *Loewe* vs. *Lawlor* case in 1908) and of labor and antitrust laws to firms, whenever they operate as labor monopsonies. Further, he claims that "antitrust law and labor law are somewhat like Siamese twins unhappily separated at birth. The natural affinity between their subject matter and concern was recognized by the early cases and statutes" (p. 50). According to the author, in the same manner that labor unions should be prohibited to exercise their monopoly power, firms should also be disciplined in their monopsony power.

Labor Laws for the Labor Market

Summing up, labor relations are based on contractual, (usually) long-term relationships, characterized by high levels of uncertainties and asymmetric information. Different to what happens in some other markets, bargaining power here is inherently uneven between suppliers (i.e., workers) and demanders (i.e., employers), and monopolies and monopsonies occur frequently. Besides that, workers' and employers' utility functions are interdependent. All this shows that market failures are abundant in this case. As economic theory tells us, regulation is needed to solve these failures. A third party – usually the government – must step in; otherwise, efficiency will not be achieved, and maximum welfare will be missed.

However, bad regulation might be worse than no regulation, and many times, the problem lays here.

The Dilemma of Labor Law in an Economic Perspective

Summing the previous discussion, one may consider the *locus* of labor as a market, but in which (legal) rules are of particular importance to equilibrate powers and interests. In this manner, both economists' and sociologists'/lawyers' perspective on labor can be equally considered, in a balanced manner. The main goal of labor law in modern democracies should be to provide sound institutions that will equilibrate information, bargaining power, and contractual relationships while fostering economic growth by promoting firms' efficiency boosting. This task should involve the *creation* of rules by the executive and legislative powers and also their *enforcement* by the judiciary. Besides that, if rules to promote social indicators (such as education, health, pension, etc.) are also developed, a country should be able to truly achieve economic development.

Yet, it seems that many countries have failed in this task, either by pending too much to one side – overregulation or "bad" regulation of the labor market, hindering potential economic growth (e.g., Latin American countries, Spain, Italy, etc.) – or by pending to the other side: lack of regulation leading to low welfare, working conditions, or inequality (e.g., Asia's developing countries and, to some degree, the USA, as compared to other industrialized countries). In some sense, this reflects a one-sided view of labor, either excessively economic or excessively based on the balance of powers.

Special Topics

In this section, we discuss the main types of regulation in some areas of labor relations. We refer to some empirical studies that try to access the impacts of these regulations in the economy.

Labor Regulation (In General)

Botero et al. (2004) analyze regulation of labor, specifically laws on employment, unionization, and collective bargaining, and social security in a sample of 85 countries. Their main result is that countries in which labor regulation is overall more invasive, have higher unemployment rates – especially of the young.

In a theoretical study, Blanchard and Giavazzi (2003) show that labor regulation creates sharp intertemporal trade-offs for workers. In the short run, workers are better off with more regulation, because their wages will be higher; yet, in the long run, regulation brings higher unemployment. In a dynamic approach – in which there are new entrant firms in the future – their model predicts that employed workers will be worse off with deregulation, both because wages will be lower and the probability of them being unemployed will be higher. On the other hand, workers who would have been unemployed without the entrance of new firms benefit either by the new possibilities of being hired or with an increase in their wages. Although this study does not include empirical observation, it seems to be a more careful analysis of the implications of regulation and deregulation in the labor market, compared to those done initially in the classical economic literature.

In the other side of the discussion – the legalistic, noneconomic literature – there have been

claims that labor regulation should be (also) approached in a transnational manner, specifically under regional integration pacts (Trubeck et al. 2000). Although the idea seems coherent, these studies would greatly benefit from a more analytical and empirical approach, which could evaluate the concrete outcomes on the overall economy, and specifically on the variables directly affecting workers and employers.

Unions

Basic economic models show that unions impact markets by raising wages of unionized workers. Consequently, employment levels decrease, since employers try to substitute those more expensive workers with nonunionized, cheaper ones. In a context where there is no perfect mobility of factors (in this case, labor is the main factor), this creates a long-lasting, perverse effect in the economy: the presence of a dual labor market (Piore 1969; Doringer and Piore 1971), in which the first one is marked by the presence of unionized, high-skilled workers, earning high wages, but in which the level of employment is lower. On the other hand, low-skilled workers are trapped in the market with lower wages and, usually, worse labor conditions. This happens because due to the lack of skills, inferior workers cannot move to the first market; on the contrary, firms tend to have higher mobility and may choose to operate in either the superior or the inferior market, depending on the presence of unions and their bargaining power. If it is the case that firms are not able to choose in which market to operate, in the long run, those facing unions will lose their comparative advantage and lose business to those who do not face unions. The result will be a decline in unionization in the overall economy (Posner 2003). In countries where unionization does not happen at the firm level, the analogy holds true for different sectors of the economy: those who face unionization will lose competitiveness and may perish in the long run; the opposite is true for those sectors not facing unionization.

For some decades, empirical literature has systematically brought evidence in this direction (e.g., Borjas 2016). Yet, one study shows these effects in a careful and detailed manner and deserves a more careful attention. Aidt and Tzannatos (2008), besides linking union activities to monetary policies in an original but important fashion, show that it is not the simple, cross-country variation in union density that affects economic outcomes (as the literature has traditionally implied). The authors bring the *coordination* and *bargaining coverages* to the spotlight. They empirically associate the existence of coordinated bargaining systems – via labor unions or other formal and informal labor organizations – to better macroeconomic outcomes (e.g., lower unemployment) and more flexible labor markets, what may sound odd at a first glimpse. On the contrary, high bargaining coverages are associated to poorer economic performance. According to these authors, bargaining coordination guarantees positive outcomes and also enables labor markets to respond more adequately and in a less adverse manner to external shocks. In an effort to employ the classical economic theory of labor, coupled with an institutional perspective, these authors conclude that "it is the total 'package' of (formal and informal) institutions that matters for economic performance" and that "labour market coordination cannot and should not be thought of in isolation from the broader institutional environment" (p. 290).

Minimum Wage

Economists usually regard the effects of mandated minimum wages as similar to those of unions, because unions' most usual demand is on higher wages. As the models predict, minimum wages create unemployment, especially for marginal workers, i.e., female, young and old, and black (Posner 2003). The effect is explained by a change in relative prices between less- and high-skilled workers, the first becoming more expensive (because minimum wages are set lower than the level of high-skilled workers' productivity, not affecting, thus, their wages). If minimum wage is applicable only to some sectors and not to others, the effect is again equal to that when there is a dual labor market, of unionized vs. nonunionized firms: it may increase unemployment in the sector covered by minimum wage and decrease unemployment in the sectors in which no mandatory wage prevails. Finally, as Posner (2003) explains, minimum

wage may hinder on-the-job training of marginal workers (those who need it the most), because employers may not compensate the costs of the training investment by paying lower wages.

This topic has brought much attention to the economic literature, and Alan Krueger calls for a "History of Economic Thought on the Minimum Wage" (2015). Traditionally, evidence brought by this literature has corroborated the classical microeconomic model, as explained above. Yet, recently some studies have questioned that explanation, by showing that the effects of minimum wages on unemployment have been overstated. Hoffman (2016), for instance, finds no impacts of increases in minimum wages on young and low-skilled workers' unemployment. In fact, the author finds some *positive* effects of wage increases on employment for a few states in the USA. The study also presents some other recent literature corroborating its findings.

Undoubtedly, although a "history of economic thought" already exists for this theme, more empirical studies, which carefully analyze the broad implications of minimum wages on economic variables, are still needed. The story is not over.

Unemployment Compensation

Traditional economic models predict that longer and more generous unemployment compensation leads to lower rates of employment and higher wages. The explanation is that, besides having less incentives to find a job, workers are comfortable to calmly search for better jobs, i.e., their outside options increase. Yet, empirical evidence coming from the literature is not strikingly convincing in this direction. By analyzing changes in the rules of unemployment compensation in the USA during the 2008–2009 recession, Nicholson et al. (2014) find that "[unemployment coverage] extensions may indeed have had detectible effects on the U.S. labor market, but some of the studies are contradictory" (p. 212). Ludsteck and Seth (2014) for Germany in the late 1990s and beginning of 2000s, and Howell and Rehm (2009) for the OECD countries in a period of 30 years, also find contradictory empirical evidence.

This is certainly a subject which deserves more in-depth empirical and analytic studies.

Outsourcing and Subcontracting

The current phenomenon of labor outsourcing takes place under two variations: cross-border outsourcing (transnational subcontracting) and within-border outsourcing (local subcontracting). The first one is a relatively recent trend, which intensified in the last three to four decades, in which companies from industrialized countries, mainly the USA and Western Europe, move their fabric plants abroad, to countries in which labor is cheaper. They may also simply stop their own national production and purchase products from other firms and factories in those poorer locations. The second type of outsourcing happens nationally and is a much older phenomenon: a firm, with its own employees, specializes in some activity – either manufacturing, assembling, or servicing – and hires other firms and/or outside workers to do activities in which it is not specialized. For instance, an automobile firm does not produce all the automotive parts, a bank may hire another firm to provide specialized security services, a movie theater may subcontract people and firms to offer popcorns, and a school may hire a company to be responsible for the cleaning services. Subcontracting, in this sense, materializes Adam Smith's idea of division of labor which occurs in a society.

As the technology advances, the economy gets more specialized, and outsourcing – either transnational or local – becomes more pervasive. Along with this process, trade unions and other advocates of labor's interests (labor sociologists, labor attorneys, public prosecutors, etc.) have loudly manifested their opposition to the increasing trends in labor outsourcing. They argue that it has been responsible for the degradation of labor conditions, including the decrease in wages and the increase of unemployment: *inside* workers lose their positions, and lower-paid jobs are created for *outside* workers (located in the same country or abroad). Once again, employers and workers seem to be in opposing sides, headed to divergent directions. What does evidence show?

Actually, very controversial results. In the topic of international outsourcing, literature is almost split: those who argue that it is beneficial for workers and others showing that it is very

deleterious to workers – for those in rich and in poor countries. Bachmann and Braun (2008) show part of the controversy in the literature. However, after analyzing a big panel of data, their conclusions are positive toward outsourcing, at least for some workers in Germany. They show that German companies cross borders to hire labor in cheaper countries, workers face higher job is stability, and the effect is more significant for those in the service sector and less in the manufacturing sector. Yet, the authors remark that effects are strongly heterogeneous across the skill levels and age. Specifically, international outsourcing is significantly hazardous for medium-skilled and lower workers in the German manufacturing sector.

Anner (2011) evaluates the impact of outsourcing on the other side of the story, i.e., in countries of *destination*, and finds evidence of very negative results. In Central America – specifically Nicaragua and El Salvador – outsourcing reduced workers' bargaining power (due to spatial dispersion) and led to lower wages.

With regard to domestic outsourcing, or subcontracting, Autor (2003) shows that unionized firms in the USA tend to outsource *more* than nonunionized firms. This is a counterintuitive result and offers hints that unions are not always able to guarantee rules that benefit labor. Yet, the author is cautious about welfare impacts. In Brazil, Stein et al. (2015) use official data on eight million workers and employ the methodology of "fixed effects panel data," i.e., they compare wage effects for a particular person when he/she migrates from being an inside worker to an outsourced one and vice versa. The main goal is to control for each worker's personal (observable and nonobservable) characteristics. Their results show that, after controlling for fixed effects, wage differential between inside and outsourced workers is of 3% only. Besides that, significant wage differentials are created when outsourcing occurs for low-skilled workers under such situations, wages are 12% lower compared to inside workers with the same characteristics.

Such controversial evidence in the literature shows the need for more detailed and careful analysis on this topic.

Other Labor Issues

Posner (2003) provides brief discussions on a long list of other labor issues, such as child and female labor, health and safety at the workplace, maternity leave, employment discrimination (race, sex, age, disability), pension law, etc. However, more empirical analysis on each of these themes is still needed, and many other themes are also still left uncovered by the literature. It is also important to directly link the explanation coming from economic models to institutional evaluations: What are the effective impacts of specific labor rules? (For instance, "What are the effects of creating quotas – based on gender, race, age, etc. – at a workplace?") What are the political, sociological, etc., concerns behind the stipulation of some labor laws? (For example, "Why should the employer be always strictly liable for damages in case of accident at the workplace? Is this the manner to minimize damage costs and probabilities of accidents? If not, what kind of rule would be more effective?")

Our discussion above offered some initial hints. But much should be assessed empirically.

Summary

We started this entry by arguing that the literature on labor relations has been very much divided, mainly between the classical economic approach and the legal/sociological one. An economic analysis of labor law would benefit by abridging both perspectives and complementing with more. It seems that the model of labor market is a plausible one. However, it is crucial to remember that, more than other usual markets, there are strong failures in place here. There are also problems of unbalanced bargaining and political powers. Because of this, labor regulation cannot mirror simple, ordinary contract regulations. A special look is needed.

Empirical literature, unfortunately, mirror the dichotomy of the economic vs. legal approach and, for several (if not all) issues, provide contradictory, inconclusive evidence. Many times, problems might have happened due to imperfect data or inadequate methodologies. As information technology advances, as well as the quality of

databases, one might expect higher quality and, hopefully, more conclusive evidence.

No matter what, although this topic is one of the oldest themes in the social sciences, there is still room for much research.

Cross-References

► Incomplete Contracts

References

Aidt TS, Tzannatos Z (2008) Trade unions, collective bargaining and macroeconomic performance: a review. Ind Relat J 39(4):258–295

Anner M (2011) Unionization and wages: evidence from the apparel export sector in Central America. Ind Labor Relat Rev 64(2):305–322

Autor DH (2003) Outsourcing at will: the contribution of unjust dismissal doctrine to the growth of employment outsourcing. J Labor Econ 21(1):1–42

Bachmann R, Braun S (2008) The impact of international outsourcing on labour market dynamics in Germany the impact of international outsourcing on labour market dynamics in Germany. Ruhr Econ Pap 58(1):1–29

Barnes P (2004) The auditor's going concern decision and types I and II errors: the Coase theorem, transaction costs, bargaining power and attempts to mislead. J Account Public Policy 23(6):415–440

Blanchard O, Giavazzi F (2003) Macroeconomic effects of regulation and deregulation in goods and labor markets. Q J Econ 118(3):879–907

Borjas GJ (2016) Labor economics, 7th edn. McGraw Hill Education, New York

Botero JC, Djankov S, Porta RL, Lopez-De-Silanes F, Shleifer A (2004) The regulation of labor. Q J Econ 119(4):1339–1382

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

Cooter R, Ulen T (2007) Law and economics, 5th edn. Addison Wesley, Boston

Doringer PB, Piore MJ (1971) Internal labor markets and manpower analysis. Heath, Lexington

Galiani S, Torrens G, Yanguas ML (2014) The political Coase theorem: experimental evidence. J Econ Behav Organ 103:17–38

Hoffman SD (2016) Are the effects of minimum wage increases always small? A reanalysis of Sabia, Burkhauser, and Hansen. Ind Labor Relat Rev 69(2):295–311

Howell DR, Rehm M (2009) Unemployment compensation and high European unemployment: a reassessment with new benefit indicators. Oxf Rev Econ Policy 25(1):60–93

Krueger AB (2015) The history of economic thought on the minimum wage. Ind Relat 54(4):533–537

Litwinski JA (2001) Regulation of labor market monopsony. Berkeley J Employ Labor Law 22:49–98

Ludsteck J, Seth S (2014) Comment on 'unemployment compensation and wages: evidence from the German Hartz reforms' by Stefan Arent and Wolfgang Nagl. Jahrb Natl Stat 234(5):635–644

Nicholson W, Needels K, Hock H (2014) Unemployment compensation during the great recession: theory and evidence. Natl Tax J 67(1):187–218

Piore MJ (1969) On-the-job training in the dual labor market. In: Arnold R, Weber FC, Woodrow LG (eds) Public-private manpower policies. Industrial Relations Research Association, University of Wisconsin., 1969, Madison, pp 101–132

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

Stein G, Zylberstajn E, Zylberstajn H (2015) Diferencial de Salários da Mão de Obra Terceirizada no Brasil. Working Paper 400, São Paulo School of Economics, FGV, CMICRO – No. 32, Working Paper Series (Ago 7). Available at: http://bibliotecadigital.fgv.br/dspace/handle/10438/13883

Trubeck DM, Mosher J, Rothstein JS (2000) Transnationalism in the regulation of labor relations: international regimes and transnational advocacy networks. Law Soc Inq 25:1187–1211

Further Reading

Aubuchon C, Bandyopadhyay S, Bhaumik SK (2012) The extent and impact of outsourcing: evidence from Germany. Fed Reserve Bank St Louis Rev 94(4):287–304

Dickens W, Franco F, Landier A, Nicoletti G, Spector D, Solow R, . . . Alesina A (2003) Macroeconomic effects of regulation and. Europe 879–908

Laporšek S (2013) Minimum wage effects on youth employment in the European Union. Appl Econ Lett 20(14):1288–1292

Economic Analysis of Law

Alain Marciano
MRE and University of Montpellier, Montpellier, France
Faculté d'Economie, Université de Montpellier and LAMETA-UMR CNRS, Montpellier, France

Abstract

The purpose of this entry is not to identify the central claims upon which rests an "economic analysis of law." That goes far beyond what

could be done. Our goal is to characterize methodologically an economic analysis of law and, as a consequence, to establish and explain the distinction that exists between an "economic analysis of law" and "law and economics."

Definition

Economic analysis of law is the field defined by the use of economics to analyze legal phenomena and the functioning of the legal system.

Introduction

The purpose of this entry is not to identify the central claims upon which rests an "economic analysis of law." That goes far beyond what could be done. Our goal is to characterize methodologically an economic analysis of law and, as a consequence, to establish and explain the distinction that exists between an "economic analysis of law" and "law and economics."

Usually, those terms are used interchangeably to describe any economic work dealing with law or legal rules. For instance, "The Problem of Social Cost" (Coase 1960) that represents the "origin [of . . .] the modern *law and economics* movement" (Hovenkamp 1990, p. 494; emphasis added) and marks the passage from an "old" to a "new" law and economics (Posner 1975) is also viewed as the article that "established the paradigm style for the *economic analysis of law*" (Manne 1993). Symmetrically, an *Economic Analysis of Law* (Posner 1973) was viewed as "coursebook in law-and-economics" (Krier 1974, p. 1697). One could also quote Cento Veljanovski (1980, p. 160) and even Richard Posner, one of the founders of an economic analysis of law, who characterized his work as one of the "recent developments in *law and economics*" (1975, emphasis added). Many additional references could be cited that would confirm that the two expressions are usually viewed as synonymous.

Yet, sometimes, they are distinguished. In this regard, Ronald Coase is probably one of the most significant authors to quote. He explained that "two parts" coexist in law and economics (1996, p. 103; or Coase in Epstein et al. 1997, p. 1138), which are "quite separate although there is a considerable overlap" (Coase 1996, p. 103). The first part corresponds to what is called law and economics and implicitly corresponds to the analyses to which Coase attached his name. The second part, to which "Judge Posner is the person who has made the greatest contribution" (Coase in Epstein et al. 1997, p. 1138), is "often called the economic analysis of law" (Coase 1996, p. 103).

This is the distinction we want to emphasize in this text. Our point is that a better understanding of an "economic analysis of law" requires careful understanding of the differences with "law and economics" and, therefore, a careful understanding of what is law and economics.

A Negative Characterization of an Economic Analysis of Law: Law and Economics

In a "law and economics" approach, the focus is put on the economy, the economic system, or economic activities, and, since economic activities take place in an institutional, legal environment, a correct understanding of the economy and of economic problems requires to take into account how and how far legal rules do affect the economy. This is precisely what law and economics is. This is exactly what Coase did in "The Problem of Social Cost (1960)." He "used the concept of transaction costs to demonstrate the way in which the legal system could affect the working of the economic system" (Coase 1988, p. 35). Later, he added: "[F]or me, 'The Problem of Social Cost' was an essay in economics. It was aimed at economists. What I wanted to do was to improve our analysis of the working of the economic system" (1993, p. 250). From this perspective, Coase was one of the founders of "law and economics" in its modern form but not of an economic analysis of law.

What is important to correctly understand the distinction between law and economics and economic analysis of law is that a law and economics approach rests on a definition of economics by scope, object, domain, or subject matter. In other words, what distinguishes economics from other social sciences is that each of these sciences has its own subject matter. Once again, this was the perspective explicitly adopted by Coase, for whom "economists do have a subject matter" (1998, p. 93). It corresponds to "certain kinds of activities" (1978, p. 206) or, more broadly, "the working of the economic system, a system in which we earn and spend our incomes" (1998, p. 93). Or, in a slightly different way, economists study "the working of the social institutions which bind together the economic system: firms, markets for goods and services, labour markets, capital markets, the banking system, international trade, and so on" (1978, pp. 206–207). In other words, economists study the activities that take place on explicit markets. That's the only set of activities that they can analyze with their tools. Coase again: economists "should use these analytical tools to study the economic system" (1998, p. 73).

This view has two implications. The first one is that economists should not study what is outside of the scope of their discipline, in particular legal rules, legal phenomena, and legal cases. They do not fall into the subject matter of economics and are not studied by law and economics. They are important but *only* to give "details of actual business practices (information largely absent in the economics literature)" (Coase 1996, p. 104). Second, law and economics does not only exclude certain objects from its domain of investigation and also excludes noneconomists from the analysis of economic activities. Coase was very clear about that: the subject matter is "the dominant factor producing the cohesive force that makes a group of scholars a recognizable profession" (p. 204), "the normal binding force of a scholarly profession" (p. 206), and what "distinguishes the economics profession" (p. 207). Thus, the delineation or delimitation of the scope of economics establishes a distinction with other social sciences and guarantees the unity and the autonomy of economics.

What Is an Economic Analysis of Law

By contrast with "law and economics" presented above, an economic analysis of law implies a radical change in the object of study. The focus is no longer put on economic activities – defined as the activities that take place on markets – and the objective is no longer to understand how legal rules influence the economy. The legal system is no longer seen as the environment in which economic activities take place and therefore external to the object of study – the working of the economic system. It becomes the object of study. In fact, and very straightforwardly, an economic analysis of law consists in using economics to analyze the legal system and how it works or, to quote Lewis Kornhauser, "Economic analysis of law applies the tools of microeconomic theory to the analysis of legal rules and institutions" (2011). This means, in particular, that legal rules are no longer taken as given and exogenous. An economic analysis of law endogenizes legal rules. To quote Posner, an economic analysis of law consists in "the application of the theories and empirical methods of economics to the central institutions of the legal system" (1975, p. 39).

From a methodological perspective, an economic analysis of law rests on and requires a specific definition of economics – completely different to the definition of economics used in law and economics; the difference in definition of economics is such that it makes law and economics and an economic analysis of law incompatible. Indeed, an economic analysis of law does not and cannot rest on a definition of economics by subject matter or by scope or by domain, as it is the case with law and economics. Analyzing the working of the legal system is possible and legitimate, only if the very idea that there exists a subject matter specific to economics, to which economists should restrict their attention, is abandoned. Otherwise, there would be no justification to analyze, among other things, the behavior of criminals, judges, prosecutors, or attorneys and any of the phenomena that are usually studied in economic analyses of law. These behaviors and phenomena are not,

strictly speaking, of economic nature because they do not take place on markets. They can be studied by economists only because it is assumed that any kind of activity and of behavior or any phenomenon, even those taking place outside of markets, can be studied by economic theory. This is exactly the view adopted and promoted by Gary Becker who stressed that economic theory "applies to both market and nonmarket decisions" (1971, p. viii) or "the economic approach is clearly not restricted to material goods and wants, nor even to the market sector" (1976, p. 6).

Let us note here that such a change in perspective – the expansion of the domain of economics beyond its "traditional" boundaries – is a consequence of the assumption that no difference exists between market and nonmarket behaviors. Individuals are supposed to *always* behave in the same way. As Becker wrote, "human behavior is not compartmentalized, sometimes based on maximizing, sometimes not, sometimes motivated by stable preferences, sometimes by volatile ones, sometimes resulting in an optimal accumulation of information, sometimes not" (1976, p. 14).

This then means that all social sciences have exactly the same subject matter. All social sciences can study the same behaviors and the same phenomena. The only difference that exists between them is the method or the approach they use (Becker 1971). Then, from such a perspective, economics is defined or rather described or characterized – Posner (1987, p. 1) argued that economics cannot be defined – by its method. Economics is a "*way* of looking at human behavior" (Becker 1993, italics added). To use Posner's words, economics is "a powerful tool" (1973, p. 3) or "an open-ended set of concepts" (Posner 1987, p. 2), which can be used to analyze any kind of human or social phenomenon – including legal ones. Then, "when used in sufficient density these concepts make a work of scholarship 'economic' regardless of its subject matter or its author's degree" (Posner 1975). It is only if this definition of economics is adopted that an economic analysis of law is possible.

A Few Historical Landmarks

Cesare Beccaria and Jeremy Bentham – sometimes Gladstone (see Posner 1976) – are viewed as the "predecessors" (Stigler 1984, p. 303) of an economic analysis of nonmarket behavior (among others: Posner 1975, 1993, p. 213) and, more specifically, of economic analyses of crime and punishment. Indeed, they were the firsts to analyze illegal behaviors and illegitimate activities as the result of an "economic calculus" (Becker 1968, p. 209) or of a "rational choice" (Posner 1993, p. 213).

In the twentieth century, the first who developed an economic analysis of law is Guido Calabresi in the early 1960s. In "Some Thoughts on Risk Distribution and the Law of Torts (1961; see also Calabresi 1965)" by contrast with Coase, Calabresi used economics to analyze a legal problem – namely, the compensation of victims of car accidents in different systems of liability. This was acknowledged by Walter Blum and Harry Kalven (1967, p. 240), Posner (1970, p. 638) and Frank Michelman (1971, p. 648). However, Calabresi also insisted that his work should not be viewed as a form of economic analysis of law. We suggested elsewhere that his analysis should be viewed as a form of heterodox economic analysis of law, mainly for two reasons. First, he rejected the behavioral assumption that economic analyses of law use – individual rationality. Second, he eventually criticized standard – read, Posnerian – economic analyses of law because it assumes that the "world" – the conditions in which individuals act and live – is given and by analyzing how individuals chose in a set of given conditions. To him, the law could be used to change the world and not to promote its economic efficiency (Kalman 2014). In other words, Calabresi claimed that economics, and economic analyses of law, should not be only about the allocation of resources. And, one could add, Calabresi did not make the methodological move of explicitly defining economics as a method.

It was Becker who did this move, because and when he was the first economist who consistently and repeatedly used economics to analyze

nonmarket behaviors and to explicitly define economics as an "approach." Among his writings, "Crime and Punishment: An Economic Approach" (1968) must be singled out as the first (modern) economic analysis of a legal problem, namely, crime and illegal behaviors. Indeed, Becker was the first to explain crime as the result of a rational decision, of an economic calculus, that is, of the comparison of costs and benefits. As mentioned above, this comes from the idea that all human behaviors are of the same nature and can be explained as if they were rational, an assumption that remains crucial to an economic analysis of law.

However, Becker's direct contributions to an economic analysis of law were scarce. The ones who really founded an economic analysis of law are William Landes, Isaac Ehrlich, and Richard Posner.

Landes and Ehrlich were Becker's PhD students (Fleury 2015, 2016). They transformed Becker's insights into a specific field of research. Ehrlich studied the participation in illegitimate activities and deterrence (1967, 1970), and Landes analyzed the effects of fair employment legislation on the well-being of discriminated nonwhites (see Fleury 2014). Also of particular importance, Landes was the first who developed an economic analysis of courts (1971). Landes's work was important because it was the first to really propose a model of the working of the judicial system, taking into account the two sides of the legal "market." Thus, while Becker had introduced the assumption that criminals are rational, Landes introduced the assumption that prosecutors also are rational.

Landes played also an important role for having involved Posner in a program in law and economics launched by the National Bureau of Economic Research (Landes 1998). Posner became one of the most important figures in economic analyses of the law. Not only he invented the expression, in the title of his 1973 book, and launched the first journal devoted to an economic analysis of law – namely, the *Journal of Legal Studies* – but he also contributed to explicitly define the field, providing its methodological bases and incessantly opening up new domains of analysis. He is one of the most important – quantitatively and qualitatively – contributors to the field. Posner's contribution cannot be described or summarized, and, accordingly, it could be said that it is particularly difficult to summarize an "economic analysis of law." However, let us note that Posner generalized the use of the assumption that individuals are rational in economic analyses of law; in particular, he developed and expanded the analyses of judicial decision making – explaining that judges make their decisions by maximizing a utility function. Actually, Posner linked this claim to another claim about the efficiency of the law, at the same time a positive and normative claim. Rationality and efficiency are two central claims, maybe the two pillars upon which rest economic analyses of law.

Conclusion

We suggest to distinguish between "law and economics" and an "economic analysis of law." An economic analysis of law is based on a definition of economics as a method, a (set of) tool(s) that can be used to analyze the functioning of the legal system. The distinction was invented in the early 1970s but remains central to understand most of the analyses made at the intersection of economics and the law.

Cross-References

- ▶ Austrian Perspectives in Law and Economics
- ▶ Austrian School of Economics
- ▶ Becker, Gary S.
- ▶ Coase, Ronald
- ▶ Coase Theorem
- ▶ Consequentialism
- ▶ Law and Economics
- ▶ Law and Economics, History of
- ▶ Posner, Richard
- ▶ Rationality

Acknowledgments Special thanks to Magdalena Malecka for her comments on a previous version.

References

Becker GS (1968) Crime and punishment: an economic approach. J Polit Econ 76(2):169–217

Becker GS (1971) Economic theory. Alfred Knopf, New York

Becker GS (1976) The economic approach to human behaviour. The Chicago University Press, Chicago

Becker GS (1993) Nobel lecture: the economic way of looking at behavior. J Polit Econ 101(3):385–409

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

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

Calabresi G (1965) The decision for accidents: an approach to non-fault allocation of costs. Harv Law Rev 78(4):713–745

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

Coase R (1978) Economics and contiguous discipline. J Leg Stud 7(2):201–211

Coase R (1988) The nature of the firm: influence. J Law Econ Org 4(1):33–47

Coase R (1993) Law and economics at Chicago. J Law Econ 36(1 Part 2):239–254

Coase R (1996) Law and economics and A. W. Brian Simpson. J Leg Stud 25(1):103–119

Coase R (1998) The new institutional economics. Am Econ Rev 88:72–74

Ehrlich I (1967) The supply of illegitimate activities. Unpublished. University of Chicago

Ehrlich I (1970) Participation in illegitimate activities: an economic analysis. PhD dissertation, Chicago

Epstein RA, Becker GS, Coase RH, Miller MH, Posner RA (1997) The roundtable discussion. Univ Chic Law Rev 64(4):1132–1165

Fleury J-B (2014) From opportunity theory to capital punishment: economists fighting crime in the 1960s and 1970s. Mimeo

Fleury J-B (2015) Massive influence with scarce contributions: the rationalizing economist Gary S. Becker, 1930–2014. Eur J Law Econ 39(1):3–9

Fleury J-B (2016) Gary Becker. In: Marciano A, Ramello G (eds) Encyclopedia of law and economics. Springer

Hovenkamp H (1990) The first great law and economics movement. Stanford Law Rev 42(4):993–1058

Kalman L (2014) Laura some thoughts on Yale and Guido. Law Contemp Probl 77(2):15–43. Available at: http://scholarship.law.duke.edu/lcp/vol77/iss2/3

Kornhauser L (2011) The economic analysis of law, Stanford Encyclopedia of Philosophy Archive. http://plato.stanford.edu/archives/spr2014/entries/legal-econanalysis/

Krier JE (1974) Review of economic analysis of law by Richard A. Posner. Univ Pa Law Rev 122(6):1664–1705

Landes W (1971) An economic analysis of the courts. J Law Econ 14(1):61–107

Landes W (1998) The art of law and economics: an autobiographical essay. Am Econ 1(Spring 1997), reprinted in Passion and Craft, Economists at Work

Manne HG (1993) An intellectual history of the George Mason University School of Law. http://www.law.gmu.edu/about/history#manne

Michelman F (1971) Pollution as a tort: a non-accidental perspective on Calabresi's costs (review of the costs of accidents: a legal and economic analysis by Guido Calabresi). Yale Law J 80(3):647–686

Posner RA (1970) The costs of accidents: a legal and economic analysis. Univ Chic Law Rev 37(3):636–648

Posner RA (1973) Economic analysis of law. Little, Brown and Co, New York

Posner RA (1975) The economic approach to law. Texas Law Rev 53(XX):757–782

Posner RA (1976) Blackstone and Bentham, The revolution in social thought. J Law Econ 19(3), 1776:569–606

Posner RA (1987) The law and economics movement. Am Econ Rev 77(1):1–13

Posner RA (1993) Gary Becker's contributions to law and economics. J Leg Stud 22(2):211–215

Stigler GJ (1984) Economics: the imperial science? Scand J Econ 86(3):301–313

Veljanovski C (1980) The economic approach to law: a critical introduction. Br J Law Soc 7(2):158–193

Economic Development

Silvia Ručinská[1], Ronny Müller[1] and Jannik A. Nauerth[2]

[1]Faculty of Public Administration, Pavol Jozef Šafárik University in Košice, Košice, Slovakia

[2]Faculty of Business and Economics, University of Technology Dresden, Dresden, Germany

Abstract

Economic development describes a process of improving economic and social well-being of people in a specific area. In addition to qualitative and quantitative growth, it includes social and environmental aspects. For this purpose various economic measures with different focus are used to quantify economic development.

Introduction

The term economic development is generally used, but there is no clear definition of it. In

general the economic conditions and the standard of living are associated with it. Especially economists tend to consider economic development in terms of economic growth. Economic growth is the percentage change in the amount of a value from one period to another. In general there is a close relation between economic development and the growth of output in an economy. Economic growth describes just a quantitative change in the level of an economic measure. Beyond this development, it describes the improvement of economic and social conditions as a hole including qualitative as well as quantitative measures. Economic growth and economic development are different categories, but for explaining the economic development is a definition of the economic growth useful.

From the Economic Growth to Economic Development

A commonly used instrument to measure economic growth is the "gross domestic product" (GDP). There are three different approaches to measure it. Considering the GDP in terms of expenditures, it is defined as "the sum of the final uses of goods and services (all uses except intermediate consumption) measured in purchasers' prices, less the value of imports of goods and services" (OECD a). The other approaches regard it in terms of income and production. Thus, it summarizes the economic power within one country into a number. Furthermore, it contains economic acting of all people within the borders of the country without taking in consideration of nationality. From this, economic growth describes the percentage change in GDP from one period to another. In general, values are reported quarterly, as well as yearly. The "gross national income" (GNI) is quite similar to the GDP, but it considers all citizens of a nation, even if they work in another country.

To refine the GDP, a per capita adjustment is used. It is calculated as the quotient of GDP and the number of residents. However, it is independent of any distributional dimension. Therefore, the GDP per capita does not refer to real distribution, it is just statistical. In relation to population, one can distinguish between extensive and intensive growth. Extensive growth describes an increase in GDP without an increase in its per capita. Contrary to this, intensive growth describes itself as an increase in the level and per capita. To compare the GDP above countries, different price levels are taken into account with purchasing power parities. This adjustment uses baskets of goods to compare the purchasing power between countries.

Despite the advantages of the GDP, it is criticized frequently for excluding social and environmental issues. It should be noted that natural disasters can have positive impacts on it. Additionally, labor without payments as voluntary work or work at home is not included in the computation. Especially illegal work takes place as an estimation. Detailed information about the critics on GDP and the difficulties of measuring development are provided by Stiglitz et al. (2009).

It is usual to describe the economic growth by the model of economic growth. One approach to model growth was made by Harrod and Domar. They were influenced by "Keynesian" economic theory and developed models based on a connection between capital stock and production volume (Domar 1946), as well as investments and aggregate demand (Harrod 1939). The "neoclassical" approach of modeling growth goes back to Solow (1956). He developed a model laying down a long run equilibrium between investments and depreciation. As soon as the optimal capital stock is reached, the per capita income will not increase any further. Additional growth can only be achieved by technological progress. With respect to the utility maximization of the households, the "Ramsey-Cass-Koopmans Model" was modeled by Ramsey (1928) and expanded by Cass (1965) and Koopmans (1965). Different to the "Solow Model," it considers the saving rate as endogenous depending on the optimization of infinitely living households. However, the technical progress is taken into account as exogenous. This critic encouraged Romer (1990) to develop a new model with endogenous technological change, depending on preferences of the households. However, the "Ramsey-Cass-Koopmans

Model" is the approach most commonly used in growth theory.

Measures of Economic Development

To describe and measure the economic development of economies on an easy way, the Word Bank offers a classification by income every year. The current values (2014) state that a GNI per capita of $1,045 per day or less describes a low-income economy, and the range from $1,045 to $12,746 describes middle-income economies. Countries with a GNI per capita and day over $12,746 are referred as high-income economies. Besides that, low- and middle-income economies are denoted as developing economies (The World Bank 2015). Another classification is used by the United Nations (UN). To define less developed countries, three conditions are used: low per capita income, human assets, and economic vulnerability (UNCTAD 2015).

Taking into consideration those critiques on the GDP and the insufficiency of quantitative indicators to measure development, it is comprehensible that many alternative measures have been developed. The UN supports the "Human Development Index" (HDI), which includes live expectancy at birth, education, and the "gross national income" per capita to measure the development more adequately. Live expectancy at birth and education are used as more qualitative indicators, even if they are numerical measures. Education is taken into account with years of schooling from educated people and the expected years of schooling of uneducated people. Live expectancy is determined by average life span at birth.

Developed countries are reported at a HDI value over 0.8. Less developed countries are reported with a HDI in a range from 0.5 to 0.8, while some sub-Sahara states reported with a value under 0.5. Besides that, the HDI is highly correlated with the GDP, which is why it is criticized as being an inadequate measure of economic development too.

Regarding the distribution of wealth, the "Gini coefficient" is commonly used. It states a value between zero and one (other scaling is possible), depending on the division of income, whereby a value of zero declares the complete equality in income and a value of one states the contrary. From that it is used as a measure of equality in a society. The OECD countries are reported with a value of 0.32 in 2012 (OECD b).

One famous critic on economic growth came from Meadows et al. (1972). This survey simulated growth scenarios considering world population, raw materials, and environmental pollution. In this context the term "qualitative growth" took place in the discussion. In contrast to "quantitative growth," it considers issues as health, education, security, and equal opportunities in addition to the GDP.

As an alternative to the GDP, Bhutan proposed the "gross national happiness" (GNH) to measure economic development. It considers social justice, environmental protection, good governance, as well as cultural values. However, it is difficult to measure those qualitative indicators and make them comparable. As the consequence, the GNH was not successful.

Moreover, critics on the GDP took place in the discussion, so the German "Federal Office of Environment" proposed the "national wealth index" (Diefenbacher et al. 2010). It includes many indicators related to environmental issues, mobility, security, health, income, and consumption.

Independent from these approaches, many measures of economic development are existing, not just considering the growth.

References

Cass D (1965) Optimum growth in an aggregate model of capital accumulation. Rev Econ Stud 32(3): 233–240

Clark G (2007) A farewell to alms – a brief history of the world. Princeton University Press, Princeton

Diefenbacher H, Zieschank R, Rodenhäuser D (2010) Measuring welfare in Germany – a suggestion for a new welfare index. Umweltbundesamt, Dessau-Roßlau

Domar ED (1946) Capital expansion, rate of growth, and employment. Econometrica 14(2):137–147

Harrod RF (1939) An essay in dynamic theory. Econ J 49(139):14–33

Koopmans TC (1965) On the concept of optimal economic growth. In: The econometric approach to development planning. North Holland, Amsterdam
Malthus TR (2007) An essay on the principle of population. Dover, Mineola
Meadows DH, Meadows DL, Randers J, Behrens WW (1972) The limits to growth. Universe Books, New York
OECD (a) Glossary of statistical terms – gross domestic product. http://stats.oecd.org/glossary/detail.asp?ID=1163. 30 June 2015
OECD (b) Compare your country – income distribution and poverty. http://www.compareyourcountry.org/inequality?cr=oecd&lg=en. 30 June 2015
Ramsey FP (1928) A mathematical theory of savings. Econ Jl 38(152):543–559
Romer PM (1990) Endogenous technological change. J Polit Econ 98(5):71–102
Solow RM (1956) A contribution to the theory of economic growth. Q J Econ 70(1):65–94
Stiglitz J, Sen A, Fitoussi J-P (2009) The measurement of economic performance and social progress revisited: reflections and overview
The World Bank (2015) Country and lending groups. http://data.worldbank.org/about/country-and-lending-groups. 30 June 2015
UNCTAD (2015) The least developed countries report 2014. United Nations Conference on Trade and Development. http://unctad.org/en/PublicationsLibrary/ldc2014_en.pdf. 30 June 2015

Economic Efficiency

Ram Singh
Department of Economics, Delhi School of Economics, University of Delhi, Delhi, India

Definition

The term Economic Efficiency refers to the relationship between aggregate benefits and costs to the individuals concerned. Among the widely used efficiency criteria are the Pareto Optimality, the Kaldor-Hicks, the Cost-Benefit, and the Wealth Maximization criterion (Hicks 1939; Jain 2015; Jain and Singh 2002; Kaldor 1939; Sen 1970 and Scitovsky 1941). In this essay, we discuss economic efficiency as a tool for the social choice among the alternative legal rules. Discussion is carried out by using illustrative examples. We show that in several contexts the efficiency can serve as a useful tool for comparing the legal rules. However, it has serious limitations as well. In several situations, the efficiency criteria can fail to compare legal rules or can lead to contradictory rankings. Moreover, the assumptions underlying some of the efficiency criteria do not hold always. We discuss merits and demerits of various efficiency criteria. It is shown that in the real world, economic efficiency has limitations as a guide for making social choice from among the legal institutions.

Introduction

The aim of the economic efficiency is to maximize the aggregate net benefits for the individuals concerned. It serves a useful tool when the objective is to maximize the net economic gains for the individuals involved. A strand of law and economics literature uses economic efficiency as a yardstick to compare outcomes under different legal rules or institutions. In other words, economic efficiency is used as a basis to compare and rank the legal rules and institutions.

The law and economics researchers use several efficiency criteria. Notable among the frequently used ones are the Pareto optimality, the Kaldor-Hicks, and the Wealth maximization criterion. In several legal contexts, these criteria can serve as a useful tool for comparing relative efficiency of the legal rules. However, on several occasions, these criteria can fail to compare various alternative legal arrangements.

For the ease of illustration, let us start with an example. Suppose there are 50 fishery units located on a water stream. Each unit earns a profit of 15 for its owner. There is plan for a cloth factory to start at upstream. If factory starts, it will generate a net profit of 600 for its owner. However, the factory will discharge chemicals in the water stream. The polluted water is going to be injurious to the fish. In the absence of any corrective measure, assume that the fisheries will suffer a harm of 10 each, that is, a total harm of 500. To keep things simple, we can assume that the factory has no other effects, good or bad, on the fisheries and also for any other third party. Moreover, assume if

a chemical treatment device is installed at the factory, it can completely solve the problem of water pollution. The device will cost 501. Now consider the following legal arrangements.

Rule 1: The factory is allowed to operate but will have to fully compensate the fishery for the loss inflicted on them – "full" compensation means that the profit of fisheries will be restored 15 each, i.e., the profit levels in the absence of the water pollution caused by the factory (see Singh 2004, 2007a). *Rule 2*: The law does not allow the factory to operate at all. Clearly, under this rule profitability of fisheries will remain unaffected. *Rule 3*: The factory is allowed to operate without any obligations to compensate the loss suffered by the fisheries. In this case, the profit enjoyed by the fishery owners will come down from 15 to 5. *Rule 4*: The factory is allowed to operate only with the treatment device installed, and half of the cost of the device is to be borne by the factory owner and the remaining half is to be split equally among the fishery owners.

Suppose the factory operates without the treatment device. Then, even after fully compensating the fisheries by paying them 10 each, the factory owner will be left with a profit of 100. Therefore, Rule 1 is more desirable than Rule 2. A shift from Rule 2 to Rule 1 makes one party better off (owner of the factory) without reducing well being of anyone else. This argument in favor of Rule 1 over 2 is at the heart of the widely used efficiency criterion called "Pareto optimality."

Pareto Optimality

According to the Pareto optimality criterion, a rule or state of affairs x is Pareto superior to another state y if and only if the following conditions are satisfied: (a) every concerned individual or party involved considers the state x to be at least as good as the state y; and (b) at least one person strictly prefers x over y. Under such a scenario, every individual will either strictly prefer x over y or will be indifferent between the two options. In our example, clearly Rule 1 is Pareto superior to Rule 2. In fact, if the factory owners were to pay 11 to each fishery owner, all parties can be made better off compared to their well being under Rule 2.

The argument can easily be extended to the social desirability of legal arrangements. A legal rule x is Pareto superior to rule y if at least one person strictly prefers x over y and everyone else is indifferent between the two. Extending the argument further, alternative x is called Pareto optimum if there is no other legal arrangement which is feasible and is Pareto superior to x. In other words if x is Pareto optimal, then there cannot be an alternative t which is available as a social choice option and is Pareto superior to x.

This point can be seen by relabeling Rules 1 in the above example as alternatives x and Rule 2 as y. Further, label Rule 3 as z and Rule 4 as w. So the set of legal alternatives which are available for social choice can be written as

$$L = \{x,y,z,w\}$$

Now, it is easy to see that legal rule x is Pareto superior to rule y. In a meaningful sense, rule x is socially more desirable than y. Therefore, the value judgment based Pareto criterion is rather appealing. If a feasible state of society say u is Pareto superior to some other state v, then plausibly u is socially more desirable than v.

From the definition, it is clear that Pareto optimality of a rule or an alternative is defined with respect to the given set of alternatives and the set of individual concerned. Any change in either of these sets can alter the optimality status of a rule or a social choice option. For instance, in our factory-fisheries example suppose for some reasons the alternative x is not available anymore. Then the social choice set will be

$$L' = \{y,z,w\}$$

where y, z, and w are the legal alternatives as defined above. Now, y becomes Pareto optimum since neither z nor w is Pareto superior to y.

Moreover, Pareto criterion can easily fail to compare the alternatives available and thereby fail to be a guide to the social choice process. To see this, consider a comparison among Rules 2 and 3, i.e., legal arrangements y and z. It can

be seen that neither *y* is Pareto superior to *z* nor the other way round. Similarly, it can be seen that legal rule *x* and *z* are not comparable on Pareto criterion. In fact, in any choice situation involving alternatives *u* and *v* if one party strictly prefers *u* over *v* and another strictly prefers *v* over *u* then the two alternatives cannot be compared, regardless of how the rest of individuals in the society feel about these two alternatives. Moreover, if the choice involves only two such alternatives, then both will be Pareto efficient!

This shows a serious shortcoming of the Pareto criterion. In real world, the social choice involves legal changes which lead to net gains to some citizens but impose net cost on the others. Specifically, consider a change in the legal status-quo such that the proposed change entails net gains to most citizens but imposes net cost on a few members of the society. Under such situations, the Pareto criterion cannot be the guide to decide between the proposed change versus the status-quo.

Kaldor Efficiency

What is known as the Kaldor efficiency criterion aims to overcome the above limitations of the Pareto criterion. According the Kaldor criterion, a shift from alternative *y* to *x* is better if the gainers can compensate the loosers and still be better-off. It is noteworthy that according to this criterion, the gainers are not required to actually compensate the losers – only requirement is that in principle they should be able to do so. On this logic, if we revisit the comparison among the Rules 2 and 3, Rule 3 is better than Rule 2 – since the gainer (factory owner) benefits by 600 out which he can fully compensate fisheries by paying 500 and still be be left with a net gain of 100. It can easily be seen that in the above example, each on of the Rules 1, 3, and 4 is better than Rule 2. In other words, alternatives *x*, *z,* and *w* are Kaldor superior to *y.* Legal alternatives *x* and *z* are equally efficient, and both are better than the 4th rule, i.e., the alternative *w.* The point is that unlike the Pareto criterion, the Kaldor criterion can compare all of the above legal alternatives. In general, the social ranking can be produced by a pair wise of comparison of the alternatives.

The Kaldor criterion is the basis of the widely used cost-benefit and wealth maximization efficiency criteria. Going back to our earlier example, a shift from rule *y* to rule *x* imposes a cost of 500 on fisheries but entails a benefit of 600 to the factory owner. Since the benefit is greater than the resultant costs, the cost benefit criterion will rank *x* better than *y.* By similar logic, rules *x*, *z,* and *w* are superior to *y.*

Wealth Maximization

The wealth maximization criterion ranks the alternatives according the levels of total wealth under them. On this count too, compared to legal rule *y*, under each one of the rules *x* and *z* the total wealth for all the parties involved is higher by 100; under *w* the corresponding figure is 99. Therefore, these rules are superior to the rule *y* according to the wealth maximization criterion. The rules *x* and *z* are all equally good. Moreover, these two alternatives are socially best, since under these rules the level of wealth is maximized.

It can be seen from the above discussion that the Kaldor, the cost-benefit, and the wealth maximization criteria provide a similar ranking of the legal rules. Moreover, the social ranking produced is complete, in that using these criteria we can compare any two legal rules. However, it should be noted that the virtues of these three efficiency criteria hold under the assumption that the wealth level is the sole determinant of individual utility levels (well beings). In the above discussion, our underlying assumption has been that well being of each party involved increases with the monetary gains enjoyed by the party.

The Kaldor criterion also serves as a useful metric when production and consumption choices involve only one good and the individual utility functions are increasing in the amount of good consumed. In that case, the criterion requires allocation resources to maximize production of the good. Once that is done, any distribution of the good among consumers is Kaldor efficient. However, such assumptions are restrictive.

Limitations of Efficiency Criteria

In real world, the consumption basket has numerous goods and services in it. Besides, the utility levels may depend on factors other than the individual consumption levels of goods and services. In several such contexts, the Kaldor criteria and its derivatives can lead to contradictory or inconclusive ranking of legal choices. Moreover, several decision makers may be guided by the non-economic considerations or might not be able to work in line with the objective of economic efficiency (see Singh 2003 and Schäfer and Singh 2018). Therefore, the real world applicability of the efficiency criteria is limited.

On top of it, the efficiency criteria completely ignore the issue of fairness, equity, and justice which can be paramount considerations when it comes to choosing from different legal regimes. Nonetheless, these criteria are widely used in law and economics. On account of the ease of working with it, most of the economic analysis of legal rules and institutions is based on the wealth maximization criterion. For a discussion on the underlying assumptions, the use, and the debates surrounding the use of this criterion see, R. A. Posner (1985). For use of the criterion in specific contexts see Singh (2007a, b) and Schäfer and Singh (2017).

Cross-References

- ▶ Cost–Benefit Analysis
- ▶ Economic Analysis of Law
- ▶ Posner, Richard
- ▶ Strict Liability Versus Negligence
- ▶ Wealth Maximization: Efficiency and Equality Considerations

References

Hicks JR (1939) The foundations of welfare economics. Econ J 49:696–712
Jain SK (2015) Economic analysis of liability rules. Springer, New Delhi, pp 23–33
Jain SK, Singh R (2002) Efficient liability rules: complete characterization. J Econ 75(2):105–124
Kaldor N (1939) Welfare propositions of economics and interpersonal comparisons of utility. Econ J 49: 549–552
Posner RA (1985) Wealth maximization revisited. Notre Dame J Law Ethics Public Policy 2:85–105
Schäfer H-B, Singh R (2018) Takings of land by self-interested governments economic analysis of eminent domain. J Law Econ 61:1–40
Sen AK (1970) Collective choice and social welfare. Holden-Day, Inc., San Francisco, Chapters 2
Singh R (2003) Efficiency of 'simple' liability rules when courts make erroneous estimation of the damage. Eur J Law Econ 16:39–58
Singh R (2004) 'Full' compensation criteria: an enquiry into relative merits. Eur J Law Econ 18:223–237
Singh R (2007a) Causation-consistent' liability, economic efficiency and the law of torts. Int Rev Law Econ 27:179–203
Singh R (2007b) Comparative causation and economic efficiency: when activity levels are constant. Rev Law Econ 3:383–406
Scitovsky T-d (1941) A note on welfare propositions in economics. Rev Econ Stud 9:77–88

Economic Growth

Maria Rosaria Alfano
Economics Department, Seconda Università degli Studi di Napoli, Italy

Abstract

Considering its quantitative property, economic growth has often been considered an index of wealth. Nonetheless, it does not represent the well-being of a given country. In fact, it lacks information on how this wealth is redistributed or on the indirect effects of the said production, such as environmental consequences. Economic literature highlights the difference between economic growth and development, attributing to the latter a holistic definition, which takes into account additional factors, such as collective well-being, social equity, life expectancy, quality of institutions, and environmental quality. Although, at a first glance, the boundaries between concepts of growth and development may be considered as clearly distinguishable, they often tend to disappear when analyzing the two variables

from a long-term perspective. In both economic theory and practice, there are several significant variables, such as the quality of human capital and institutions or environmental efficiency, which play a key role in the opportunities for development in most high-income countries. Within this framework, development becomes a fundamental feature of economic growth, when the latter is interpreted as a general improvement in quality of life, as opposed to a mere quantitative increment in production.

Definition

Economic growth is the increased capacity of an economy to produce goods and services, comparing one period of time to another. It is measured through the growth rate of the gross domestic product (or gross national product, or national income), which is calculated in real terms in order to obtain an indicator that is not influenced by inflation.

Economic Growth

Three main approaches are attributed to economic growth theory:

- Classical economists – with authors such as Smith (1776), Malthus (1798), Ricardo (1817), Ramsey (1928), Young (1928), Schumpeter (1934), and Knight (1944).
- Keynesian and neoclassical economists – who can be divided into two categories, characterized by the different role that technology plays in the production function. On the one hand, Solow (1956) and Swan (1956), Cass (1965), Koopmans (1965) regarded technology as a given variable and built models of what is referred to as exogenous growth. On the other hand, Keynesian authors, such as Harrod (1939) and Domar (1946), as well as some neoclassical authors, such as Romer (1986), Lucas (1988), Rebelo (1991), Grossman and Helpman (1991), and Aghion and Howitt (1992, 1998), directly included technology variables into the production function, implementing models of what is referred to as endogenous growth.
- Modern economists – who are all the economists interested in the immaterial aspect of growth and in the role of institutions in the economy. The most important authors in this category are North (1990), Evans (1995), Coleman (1990), Putnam (1993), Fukuyama (1995), and Sen (1999).

A different classification of economic growth theories is based on the perspective of divergence or convergence of the economies of different nations (de la Fuente 2000). Conventionally, convergence is considered when poorer economies grow at higher rates than richer ones, thus reducing income differences between them. Subsequently, divergence is determined by the opposite mechanism. According to the former, a general equalization of income levels would be observed in the long run, while an increase in the existing gap is expected when considering the latter. Indubitably, the source of such discrepancies lies in the differences in the formulation of the production function and of technological progress dynamics. Specifically, one of the conditions for economic convergence is the presence of decreasing returns to scale, which allows poorer economies, with initially lower levels of capital, to grow faster than the richer ones. Similarly, increasing returns to scale bring about divergence between the economies of different nations. Another element that characterizes the long-run growth path is related to technological progress. Though widely accepted that an increase in technological investment ensures higher future growth rates, it is not clear whether this would ultimately lead to a convergence or divergence effect. In fact, since significant differences in growth rates are not sustainable, if the technological return is a decreasing function of its accumulation, equalization in the level of technical efficiency will be observed. Finally, another factor of convergence is associated to the allocation mechanism that allows focusing investment where productivity and return on capital are greatest. Since investment

in poorer countries is concentrated in the agricultural sector, which is characterized by low labor productivity, the reallocation of resources towards the manufacturing sector allows for a rapid increase in mean productivity. Such a boost is not feasible in rich countries, where investment is already focused in highly productive sectors.

Classical Economists

When individually analyzing the three theoretical approaches to economic growth, one can immediately recognize that, aside from the main basic concepts, the classical literature contains some crucial perspectives that would later be reconsidered in modern theory. In his book *An Inquiry into the Nature and Causes of the Wealth of Nations*, Smith, founder of modern economic science, considers a nation's growth as the result of an enhancement in productive processes, which delivers an increase in wealth and production with the same resources. The latter is made possible by the augmentation of labor specialization, given by an increase in production and knowledge. Thus, the issue of low resources is overcome through enhanced efficiency. Continuous growth in the economy is thereby guaranteed.

Ricardo, on the other hand, expresses a different view on the matter, focusing on the problem of income distribution and its impact on economic development. Motivated by the conviction that development depends on the accumulation of capital, in other words, the share of profits devoted to investment, Ricardo dedicated his work to the study of the determinants of capital accumulation. With reference to an economy composed of landowners, capitalists, and workers, though no technical progress, the author showed how an increase in production could only be achieved through a greater exploitation of the employed labor force and cultivated land. According to his analysis, this process leads to a rise in rents and a continuous reduction in profits (due to the diminishing returns on land), thus discouraging investment and hampering growth. The economist Thomas Malthus, who studied the problem of long-term growth from a demographic perspective, also reached a similar conclusion. In this case, one of the key hypotheses is the absence of technological progress. The main assumption behind Malthus' work is the existence of a positive relationship between the wealth of a nation and its birth rate. Specifically, he argued that any wage level in excess of the subsistence one would incentivize individuals to expand their families, thus resulting in an increased consumption of resources for the livelihood of the population and a reduction in the proportion devoted to the accumulation of capital. Therefore, for Malthus as for Ricardo, the long run is inevitably characterized by stagnation of the economy, unless (as in China) policies directed towards controlling the level of births are implemented.

Technology is once again the fulcrum and engine of economic growth in Schumpeter's theory. The latter emphasized the entrepreneur as being the heart of the innovation process that allows the generation of short-term monopoly profits. The guarantee of compliance to rights (patents) protects the investment made by firms in the field of research. Like Smith, Schumpeter also argued that the accumulation of knowledge and the implementation of human capital produce increasing returns in the long run.

Nonetheless, this model is differentiated by the fact that entrepreneurs see innovation as the primary goal of their activity, because competition between firms is entirely based on the destruction of the monopolistic position of competitors and the acquisition of a market niche. This process, defined as "creative destruction," brings the economy into a market characterized by monopolistic competition, where increasing returns are the consequence of continuous increments in production technology and human capital. The State's role in facilitating the process of economic growth becomes crucial therefore it is based on the implementation of policies capable of creating the most fertile conditions to allow businesses to grow and create knowledge and innovation in a specific area. The said policies are economically justified when considering how avoiding less than optimal production becomes the goal of public intervention. The explicit introduction of technology in the production function also allows for the comparison of this model to those of endogenous growth.

Keynesian and Neoclassical

The Great Depression gave rise to a new generation of economists, who set out to reinterpret and reconsider the main features and determinants of economic growth, through the application of new formulations of the production function. Working within a Keynesian framework, Harrod (1939, 1942), and Domar (1946) were the first to highlight the instability of the capitalist system, given by the low substitutability between factors of production. Nonetheless, this hypothesis was overcome by Solow (1956, 1957) and Swan, who used a different production function to develop a model that had diametrically opposite consequences on economic growth. According to the latter, in the absence of technological improvements, per capita growth would tend towards a steady state (Uzawa 1965). The consequent assumption of the said model is that of convergence among different countries (Barro and Sala i Martin 1992, 1997). Subsequently, Romer (1994) and Lucas (1988) developed an analysis that took into account reproducible factors within the production function. The assumption of diminishing returns to capital was thereby overcome and endogenous growth models started to be developed.

In order to better grasp the theoretical journey described above, let us consider a generic production function with two factors of production, capital and labor: $Y_t = F[K_t, L_t, t]$, where Y_t is the production level at time t. In this expression, time takes on a role of proxy for technological development, given that, assuming equality of capital and labor, there will be a higher production level at time $t + 1$ compared to time t. The increase in capital stock at any given moment will thus be given by $\partial K/\partial t = I - \delta K = s\,F(K, L, t) - \delta K$, where δ is the depreciation of capital and I the level of investment, expressed in the latter part of the equation as the savings rate ($0 < s > 1$) multiplied by the abovementioned production function.

Harrod and Domar's model presents a peculiar formulation of the production function. The latter is that put forward by Leontief (1941); in other words it consists in fixed coefficients that could be expressed as $Y = F(K, L) = \min(AK, BL)$, with A and B being positive constants. The equilibrium condition given by the full employment of factors of production is achieved at the point where $AK = BL$, whereas when the quantities of labor and capital invested are not at equilibrium, there will be cases where $AK > BL$ or $AK < BL$. In the first case scenario, only an amount of capital $(B/A)L$ would be exploited; just as in the second case scenario, it would only be an amount of labor $(A/B)K$. These two situations would imply, respectively, an underuse of capital and labor. The hypothesis of non-substitutability implied in this production function translates into a process of continued instability, as a result of the typically Keynesian animal spirit reasoning. Thus, when starting from an ideal condition of general equilibrium, where the economy's growth rate (g) is equal to the growth rate in production (g*), there would be a surplus of supply together with a recessive spiral any time that $g < g^*$ and, vice versa, an excess in demand with inflationary expansion whenever $g > g^*$. This process would inevitably result in an economic growth determined by macroeconomic unbalance and lacking any sort of natural force that could allow for long-term equilibrium.

With the aim of overcoming the limitation posed by the non-substitutability of factors of production, the Solow-Swan model utilizes a typically neoclassical production function. This model is characterized, first of all, by the absence of a technological development variable and thus of the properties typically attributed to functions that do include the latter: namely, the factors of production have a decreasing marginal productivity, the function presents constant returns to scale, and the marginal product of capital (or labor) is infinite for K (or L) tending to zero and zero in the opposite case scenario. Through the application of this specific production function, it is possible to reformulate the equation that explains the variation of capital stock over time as $\partial k/\partial t = s\,f(k) - (n + \delta)\,k$, where all variables are expressed in per capita terms and (n) represents the population's growth rate. This equation highlights that the capital's possible evolutionary routes would be:

- $s \cdot f(k) > (n + \delta) \cdot k$: arginal increase in capital over time
- $s \cdot f(k) < (n + \delta) \cdot k$: marginal decrease in capital over time
- $s \cdot f(k) = (n + \delta) \cdot k$: constant growth in capital, in other words, characterized by the per capita values (k, y, c), which remain constant over time, and the levels of these same variables (K, Y, C), which grow at the same rate as the population (n)

In Solow and Swan's envisioning, the first two cases are unstable and converge towards the third, which represents the only long-term equilibrium possible. This is instantaneously apprehended if s f(k) is interpreted as the savings quota destined to investment and $(n + \delta)$ k as the real depreciation of capital. Net investment will always be positive in the first case and negative in the second, whereas there will be no incentives to changing the level of capital in the third case scenario. Convergence towards the steady state is inevitable, determining in the long term an absolute condition of equality between countries de la Fuente (2000). According to the authors, the countries starting with similar conditions will immediately converge to the same steady state, while the ones starting off as more disadvantaged will grow at a faster rate, to then also converge. This framework, known as absolute convergence, has been criticized by Barro and Sala i Martin (1992, 1997, 2004), who sustained conditional convergence. The latter differs from the former in that it takes into account structural differences between countries.

The decreasing marginal productivity of factors of production plays a fundamental role in ensuring the process of convergence (Baumol 1986). During the 1980s, the growing international integration, together with the need to take into consideration variables that could capsize this hypothesis, led many economists to study and develop new models that could identify growth paths from within the production function. A simplified formulation of the said models is given by the equation $Y = AK$, where A represents a positive constant for the level of technology. In comparison to the production function developed by Solow-Swan, this one is characterized by the absence of decreasing returns given by the presence of non-reproducible factors in K, particularly human capital. The latter's increasing returns counterbalanced the decreasing returns to capital, resulting in complex constant returns for each individual enterprise. When considering these new assumptions, convergence is no longer guaranteed, be it absolute or conditional. Other approaches aimed at eliminating decreasing returns to capital are based on the concept of learning by doing, introduced by Arrow (1962). The latter consists in the possibility of implementing productivity simply according to the experience acquired during previous production cycles (Sheshinski 1967). Starting from this intuition, first Romer and then Lucas implemented growth models where the accumulation of physical and human capital still guarantees constant returns for the enterprise; however, the positive externalities given by spillover effects result in increasing returns for the overall sector. In Romer's model, the (Cobb-Douglas 1928) production function becomes $Y = AK^{\alpha}(KL)^{1-\alpha}$, where α represents the quota of national per capita income destined to capital. The growth rate is constant, and this results in a permanent condition of steady state for the economy. Moreover, while an increase in knowledge will determine an immediate rise in the economy's growth rate, the boost of the savings quota destined to investment will permanently influence growth. Also in this case scenario, the investment in learning and new knowledge acquired by the entrepreneur will create positive externalities for the entire sector through the spillover effect, thus justifying policy interventions supporting innovation on the part of the State.

Human capital then assumes a central and relevant role in Lucas' model, where the emphasis is on the role of education as the only sector really capable of producing innovation. In his model, workers, for every unit of time, are asked to choose between work and education. Clearly, with the former option, they would benefit from a higher level of current income, whereas with the latter they would be able to improve their skills, thus obtaining higher productivity and income in

the future. Having defined H as total human capital, and h as the knowledge of a single worker, the following relation may be formulated: $H = hL$; if (u) represents the time a worker dedicates to the productive process, (1-u) will subsequently be the quantity of time dedicated to education and training, so that the equation determining the accumulation of human capital is $H = H\phi H(1 - u)$, where ϕ is a scale parameter, whereas the production function including human capital is given by $Y = AK^{\alpha}(uH)^{1-\alpha}$. From these equations it can be concluded that human capital is, in particular, a function of the time dedicated to education and training and that its marginal productivity is constant. In terms of general macroeconomic equilibrium, it will be found that, once the steady state is achieved, the stock of physical capital, like production and other related variables in the model, will continue to rise at a rate equal to the endogenous growth in human capital.

The Role of the Public Sector in Economic Growth Models

In neoclassical growth models, the public sector holds a role of agent responsible for the exogenous stimulation of investment and the education and training of human capital. Barro (1990) built a growth model where the goods and services produced by the public sector are considered as production inputs for the public sector. In his model, goods are divided between three main categories:

- Private, rival, and excludable goods, produced by the public sector
- Public, non-rival, and non-excludable goods, produced by the public sector
- Goods produced by the public sector and subject to congestion, i.e., rival non-excludable goods, such as roads, water supply, and sewers

In the first model, where G refers to the public sector's aggregate expenditure, the very nature of the (private) goods that are produced results in the quantity of available inputs being equal to $g = G/n$, where n represents the total number of producers. The introduction of the public sector in the Cobb-Douglas (1928) production function results in the following equation: $y = Ak^{1-\alpha}g^{\alpha}$, where returns of capital are assumed to be decreasing and the level of public expenditure to be fixed. Each agent adopts a production technology that combines capital and public goods provided by the State, where y, k, and g, respectively, represent product, physical capital, and productive public investment per worker. A specific assumption is that entrepreneurs, given a fixed level of public expenditure, will determine the quantity of private inputs (k) that will be employed.

If the government is looking to maintain a neutral balance, one potential financing methodology could be the introduction of a tax proportional to the quantity of outputs equal to $\tau = g/y$. The latter allows us to derive the condition of efficiency determined by a given quantity of goods produced by the public sector. In fact, since each goods unit requires an increased use of resources, the natural condition of efficiency will be given by $\partial y/\partial g = 1$, which would in turn imply $g/y = \alpha$. Under these assumptions, Barro and Sala i Martin (1992, 1997, 2004) demonstrates that the profitability of investment is independent of the economy's growth rate, which will instead be influenced by the quantity of services produced by the public sector. Moreover, with a constant τ (og/y), the model does not allow for periods of transition, but rather gives a growth rate equal, in each period, to the steady state's growth rate.

In conclusion, it is interesting to analyze the possible scenarios determined by the government managing to produce efficiently, thus respecting the condition $g/y = \alpha$. If marginal taxation (τ) is equal to zero, returns on investment, be it private or social, will be identical. However, in the case where $\tau > 0$, private returns will be lower, and a condition of Pareto efficiency may be achieved by employing a lump-sum form of taxation on consumption or through the introduction of a subsidy to the purchase of capital goods on the part of the government.

On the other hand, in the case where the public sector produces public goods as defined by Samuelson (1954), the production function is modified according to the introduction of the hypothesis of

non-rivalry in consumption. Consequently, the whole quantity G of each producer will be presented as opposed to the relative per capita quota, resulting in $y = Ak^{1-\alpha}G^{\alpha}$. Non-rivalry implies that the marginal product of public goods will be given by the effect of the variation in G on aggregate output Y = y n and the corresponding condition of efficiency will be given at the point where $G/Y = \alpha$. Having taken into account these minor differences, the model maintains all the implications already presented in the case of the production of private, rival, and excludable goods.

However, different implications are given when the public sector produces goods that are subjected to congestion, in other words, being characterized by rivalry and non-excludability in consumption. The phenomenon of congestion derives from the difficulty of excluding enterprises from the consumption of a given rival good, which would thus be exhaustible. Thus, for a given level of production (G) of a good, each enterprise will see a decrease in the quantity available to it, as other enterprises increase their consumption of the said good. Consequently, the production function for each producer will include a relation between the total quantity of product G and the aggregate quantity of private inputs K; therefore, $y = Ak\ (G/K)^{\alpha}$. Since the increase in capital (k) and in production (y) of a given enterprise congests the inputs (g) available to the remaining enterprises, in the absence of a proportional tax on outputs or on income, there will be an excessive consumption of "public" goods. The introduction of a tax equal to $\tau = G/Y$ will, on the other hand, be able to match the rate of returns on social and private investment, thus bringing about a Pareto optimal growth rate.

Modern Thought

While an increasing number of academics concentrated on formalizing and arguing in ever greater detail the role of human capital, education, technology, and innovation in economic growth, a group of economists, motivated by the certainty that the abovementioned elements do not constitute the engines of growth, but rather represent the consequences of the latter, set out to determine new variables that would result in being essential for development. The said research brought them to the identification of a series of elements that may be defined as immaterial and consist in the role of institutions, regulations, faith, and cooperation within society.

Institutions and Economic Growth

According to Acemoglu (2008), institutions influence the growth trajectory. The first approach to this facet is attributable to North, who defined the concept of institutions as the rules of the game within a society or, more formally, the limits conceived by man to determine human interaction. First and foremost, it must be noted how North's definition of institutions moved away from that generally conceived in everyday language. Entities such as parliament, enterprises, universities, or associations are defined by the author as organizations and are thus considered agents for change. The key concept used as a determinant of growth is, in fact, the relationship and interaction between institutions and organizations. Since the latter are born and die as a response to the set of incentives and limits imposed by institutions, the understanding of the nature of the latter, as well as how they change, represents a key for the historical evolution of countries. When analyzing the role of institutions within the economic system, the level of cooperation is identified as being a fundamental element. By applying game theory to a neoclassical structure with limited rationality, North understood that the role of institutions consisted in the highest possible reduction of transaction costs in the exchanges between agents, with the aim of rendering them convenient and encouraging their diffusion. It is clear that the level of cooperation is inversely correlated to the number of existing institutions, since when there is knowledge of the different parts and of repeated exchange (high level of cooperation), each agent does not need to feel protected by formal regulations and thus does not consider the presence of institutions as necessary. On the contrary, when there are impersonal exchanges and a high risk of disloyal behaviors, institutions acquire a fundamental role in the reduction of information costs and in the control and implementation of agreements. In conclusion, it can be stated that

the quality of regulations (institutions) influences the level of production, through the diffusion of exchanges, and subsequently affects economic growth.

Another approach, which may be considered similar to North's, however more centered on the role of the State in the promotion of an efficient economic system, is that implied by Evans in his conception of the embedded autonomy. The latter describes an ideal condition that governs the relationship between the State and civil society. In "embedded" States, one would find public officials, and institutions in general, that are deeply intertwined with economic actors and dedicated to trying to understand and favor the interests of the latter. A rooted territoriality allows for the implementation of "fitted" policies aimed at fulfilling the competitive needs of enterprises, thus favoring their development and international competitiveness. Obviously, an excessive degree of embeddedness brings about the risk of a loss of authority and of impartiality. For this reason, the State must preserve its autonomy when the moment comes to take decisions, trying to avoid being a mere instrument for the satisfaction of individualistic interests and always preserving the collective well-being as its foremost objective. To sum up Evans reasoning, a simple comparative advantage in resources of a given country will not necessarily imply a success in the corresponding sector, but will rather depend on the functioning of the existing social and political institutions. If a wealth in resources is accompanied by a predatory State, which incentivizes renters and allows for widespread lobbying, thus losing the required autonomy in decision making, then development is unlikely to be achieved. On the other hand, a State that implements development policies and is characterized by the incentivizing of entrepreneurs, meritocratic recruitment, dedication, cohesion, and social loyalty will promote efficient and profitable investments on the part of enterprises, which will thus result in long-term growth and development.

Corruption and Economic Growth

The emphasis applied by North on the role of regulations in economic growth stimulated for many economists an interest towards understanding how bureaucratic machines, with all their dishonesty and obstructive presence, are able to slow down the process through which technological progress transforms into new tools and productive processes.

Many theoretical and empirical articles in economic, social, and political literature have studied how corruption affects economic development. Their authors have mainly concentrated on the relationship between corruption and economic growth, not always finding coherent results. On the one hand, Leff (1964) and Huntington (1968) argued that corruption could be positively correlated to economic performance in the presence of a thick and cumbersome bureaucracy. According to their reasoning, bribery may allow firms to get things done, thus increasing their efficiency and enhancing economic growth. Subsequently, corruption could be considered as growth enhancing in that it acts as a lubricant within a rigid bureaucracy. On the other hand, corruption can be seen as a type of government inefficiency, since it discourages investments, due to the wide discretionary power of public officials, as well as reducing the quality of public infrastructure and services, decreasing tax revenue, and affecting the allocation of entrepreneurial skills, thus slowing down economic growth (Bardhan 1997; Mauro 1995, 1998). Specifically, corruption diverts the public budget destined to social services, securities, and education, health, and general services. This implies inefficient behavior on the part of the government, in its implementation of policies that are not always in the country's best interest (Powell 2004). The market allocation of resources is thus distorted, negatively influencing investments on the part of economic agents, reducing the quality of public infrastructure and services, and thus hindering economic growth (McMullan 1961; Tanzi and Dawoodi 1997; Mauro 1995). In addition, corruption affects both the total regional amount of public spending and its structure, directing expenditures towards sectors where bribes are easier to collect. In this respect, the econometric results found by Del Monte and Papagni (2007) on the Italian regional dataset "show two distinct negative effects of

corruption on economic growth. One effect seems to be that on private investment; the other is on the efficiency of expenditures on the part of public investment." It is therefore derived that "policies to deter corruption and to increase the efficiency of local public institutions could give very positive impulses to economic growth" for the development of southern Italy.

Social Capital and Economic Growth

Within the field of study related to the immaterial aspects of economic growth, the main contributions are undoubtedly given by J. Coleman and R. Putnam, both of whom consider social capital as being an engine of growth. Despite the lack of an unequivocal and widely accepted definition of social capital, it can certainly be distinguished from the concept of human capital, given that it is related to the set of social relations available to a subject or a group of subjects. At the core of this definition is the relational element. From the latter, in fact, one can derive the definitions of three different conceptions of social capital. The first one links social capital to the idea of members of a collective adhering to shared regulations; the second identifies it as the ability of a single subject to activate and manage interpersonal relations; and the third one associates it to the individual's capability to create networks that are useful for the achievement of one's own objectives. From these definitions it can be inferred that social capital and institutions are two inversely correlated subjects. The higher the presence of social capital, the lower will be the need for institutions in terms of guarantor for the implementation of agreements, since social sanctions will play a more prominent role. Within this framework, Coleman's approach is based on the neoclassical assumption of full rationality, however considering each economic agent, not individually, but as a part of a network of relationships. In the latter, there will be certain leading individuals that will impose their own rules; thus, the social relations between the subjects constituting the network will be the only leverage they will have to condition the approval and implementation of norms in their favor.

Putnam's work, on the other hand, can be traced back to the function of social capital as a determinant of the difference in growth between countries (Alfano and Baraldi 2012). In other words, he is interested in the analysis of the existence of loyalty and of ethical norms of reciprocity and cooperation, as a tool for analyzing the role that interpersonal relations can play in different countries' experiences of growth.

The most important empirical contribution made by Putnam is summarized in his work "Making Democracy" (1993), which develops a comparative study of Italian regions, attributing the divergence in institutional and economic performance between the North and the South to the differences in their relative endowment of what he refers to as social capital. In fact, northern Italy developed faster than southern Italy because the former has been better endowed with social capital. Considering "a region's chance of achieving socioeconomic development during this century has depended less on its initial socioeconomic endowments than on its civic endowments, [the] correlation between civics and economics reflects primarily on the impact of civics on economics, and not the reverse" (Putnam 1993, p. 157). In conclusion, Putnam takes into consideration a historical path dependency, according to which "social patterns, plainly traceable from early medieval Italy to today, turn out to be decisive in explaining why, on the verge of the twenty-first century, some communities are better able than others to manage collective life and sustain effective institutions" (Putnam 1993, p. 121) that promote economic growth.

Within this same field, Fukuyama's study on trust has contributed to the increased attention to the relevance of social capital in economic growth. According to Fukuyama (1995), societies endowed with generalized trust enjoy a form of social capital that, together with traditional factor endowments such as labor and capital, contributes to their success in modern economic competition. From this conception of the role of loyalty, it can be observed that countries with an elevated endowment of social capital are differentiated by the presence of a great enterprise, while those with only a low level of social capital are characterized by a productive structure dominated by small enterprises, where both ownership and

management are family based. The author thus marks a red line between social capital, family, loyalty, and a country's ability to generate capital and utilize it in large-scale investments. In a society with little faith in those outside the family, one would encounter what is known as familism, a phenomenon that leads the family to close itself off, thus hindering relations with the wider community. The said community may be viewed as the perfect equilibrium between the excessive isolation of the family on the one hand and the disproportionate presence of the State on the other, which discourages action on the part of spontaneous groups. According to the author, it is between these two identities that social capital and trust based on shared values are built, thus creating the expectation of a correct behavior and allowing for the birth of great organizations with evolved systems of management control.

Last but not least, reference should be made to the economist Amartya Sen, who, in his work *Development as freedom*, highlights the inappropriateness of the indicator of economic growth as a proxy for the real well-being and economic progress in a society, thus providing the decisive impulse towards its abandonment, in favor of new measures of development. He highlights how economic growth is closely connected to a function of social well-being that is highly utilitarian and that does not take into account the distribution of wealth but only its total amount. By relegating the concept of economic growth to such an index, there is the risk of paradoxically considering as blossoming economies, those found in countries where there is a rapid increase in production on the one hand and a complete absence of other important elements on the other, such as freedom and life expectancy, the opportunity to escape avoidable illnesses, the possibility of finding one's desired employment, and to live in a peaceful society free from crime.

The importance of these elements in the economic growth of countries has become clear and widely accepted among economists. Even a portion of the current entrepreneurial class seems to have understood the importance of adopting policies that allow for corporate growth in a more socially friendly manner. The latter can explain the ever-increasing implementation of corporate social responsibility (CSR) policies and social balances that highlight ethical behavior and attention to social issues on the part of the enterprise, with the aim of developing a more beneficial image both at local and at international level, thus subsequently generating higher profits. The application of this concept in balances and national policy strategies represents, without a doubt, the current frontier of the concept of economic growth. Indeed, in July 2010, the European Commission officially committed to a new policy towards sustainable and equitable development, centered on the very concept of CSR. The said policy has been named Europe 2020 Strategy and consists in a series of interventions for long-term economic growth, social cohesion, and environmental safeguard. The financial crises that started in 2008 highlighted the inappropriateness of the growth models that had thus far been followed by most states. With the aim of closing the gaps found in the said growth models and creating the preconditions for a different, smarter, more sustainable, and equitable type of development, a document has been developed describing key objectives to be achieved by 2020. The latter may be summarized as follows: with regard to labor, the goal is to increase by 75% the rate of employment for those aged between 20 and 64 years old; for research and innovation, the aim is to increase by 3% of the EU's GDP the investments in research and development; as for the safeguarding of the environment, the objective is to reduce greenhouse gas emissions by 20% compared to 1990, as well as increasing by 20% the supply of energy coming from renewable resources; with regard to education, the target is the reduction in school dropout rates by 10% and to increase by 40% the 30–34 year olds with higher education; finally, in the fight against poverty and marginalization, the European goal is to reduce by 20 million the number of inhabitants at risk of, or already living in, poverty. It is interesting to point out that current European policies on growth still contain a great portion of the economic thought analyzed above, specifically when it comes to theories on human capital, on the importance of research and development, and

on social responsibility. It should thus be noted that there is still a significant lack in policies aimed at developing social capital, and improving the quality of institutions, and that only in 2010 actions were taken towards the modification of growth models according to economic theories that had been formulated and proven years before. This institutional delay in embracing the suggestions transmitted by academics and researchers should, therefore, be taken into account as an important variable in future developments of models of economic growth.

Acknowledgement I thank Marco Bottone for helpful comment

References

Acemoglu D (2008) Introduction to modern economic growth. Princeton University Press, Princeton

Aghion P, Howitt P (1992) A Model of Growth Through Creative Destruction. Econometrica 60(2): 323–351

Aghion P, Howitt P (1998) Endogenous growth theory. MIT Press, Cambridge, MA

Alfano MR, Baraldi AL (2012) Is there an optimal level of political competition in terms of economic growth? Evidence from Italy. Eur J Law Econ. https://doi.org/10.1007/s10657-012-9340-5

Arrow KJ (1962) The economic implications of learning by doing. Rev Econ Stud 29:155–173

Bardhan P (1997) Corruption and development: a review of issues. J Econ Lit XXXV:1320–1346

Barro RJ (1990) Government spending in a simple model of endogenous growth. J Polit Econ 98(5):S103–S125

Barro RJ, Sala-i-Martin X (1992) Public finance in models of economic growth. Rev Econ Stud 59(4):645–661

Barro RJ, Sala-i-Martin X (1997) Technological diffusion, convergence, and growth. J Econ Growth 2(1):1–26

Barro RJ, Sala-i-Martin X (2004) Economic growth, 2nd edn. MIT Press, Cambridge, MA

Baumol WJ (1986) Productivity growth, convergence, and welfare: what the long-run data show. Am Econ Rev 76(5):1072–1085

Cass D (1965) Optimum growth in an aggregative model of capital accumulation. Rev Econ Stud 32(3):233–240

Cobb CW, Douglas PH (1928) A theory of production. Am Econ Assoc (Suppl) 18(1):139–165

Coleman JS (1990) Foundations of social theory. Harvard University Press, Cambridge, MA

de la Fuente A (2000) Convergence across countries and regions: theory and empirics. EIB Pap 5(2):25–45

Del Monte A, Papagni E (2007) The determinants of corruption in Italy: regional panel data analysis. Eur J Polit Econ 23(2):379–396

Domar ED (1946) Capital expansion, rate of growth, and employment. Econometrica 14(2):137–147

Evans P (1995) Embedded autonomy. Princeton University Press, Princeton

Fukuyama F (1995) Trust: the social virtues and the creation of prosperity. Free Press, New York

Grossman GM, Helpman E (1991) Innovation and growth in the global economy. MIT Press, Cambridge, MA

Harrod RF (1939) An essay in dynamic theory. Econ J 49(193):14–33

Harrod RF (1942) Toward a dynamic economics: some recent developments of economic theory and their application to policy. Macmillan, London

Huntington SP (1968) Political order in changing societies. Yale University Press, New Haven

Knight FH (1944) Diminishing returns from investment. J Polit Econ 52(3):26–47

Koopmans TC (1965) On the concept of optimal economic growth. In: McNally R (ed) The econometric approach to development planning. North Holland, Amsterdam

Leff N (1964) Economic development through bureaucratic corruption. Am Behav Sci 8(3):8–14

Leontief WW (1941) The structure of the American economy, 1919–1929. Harvard University Press, Cambridge

Lucas RE (1988) On the mechanics of economic development. J Monetary Econ 22:3–42

Malthus TR (1798) An essay on the principle of population. W. Pickering, London, 1986

Mauro P (1995) Corruption and growth. Q J Econ 110(3):681–712

Mauro P (1998) Corruption and the composition of government expenditure. J Public Econ 69(2):263–279

McMullan M (1961) Theory of corruption. Sociol Rev 9(2):181–201

North DC (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge

Powell BG (2004) The chain of responsiveness. J Democracy 15(4):91–105

Putnam R (1993) Making democracy work: civic tradition in modern Italy. Princeton University Press, Princeton

Ramsey F (1928) A mathematical theory of saving. Econ J 38(152):543–559

Rebelo S (1991) Long-run policy analysis and long-run growth. J Polit Econ 99(3):500–521

Ricardo D (1817) On the principles of political economy and taxation. Cambridge University Press, Cambridge, 1951

Romer PM (1986) Increasing returns and long-run growth. J Polit Econ 94(5):1002–1037

Romer PM (1994) The origins of endogenous growth. J Econ Perspect 8(1):3–22

Samuelson PA (1954) The pure theory of public expenditure. Rev Econ Stat 36(4):387–389

Schumpeter JA (1934) The theory of economic development. Harvard University Press, Cambridge, MA

Sen A (1999) Development as freedom. Oxford University Press, Oxford

Sheshinski E (1967) Optimal accumulation with learning by doing. In: Shell K (ed) Essays on the theory of optimal economic growth. Massachusetts Institute of Technology, MIT Press, Cambridge, pp 31–52
Smith A (1776) An inquiry into the nature and causes of the wealth of nations. Random House, New York, 1937
Solow RM (1956) A contribution to the theory of economic growth. Q J Econ 70(1):65–94
Solow RM (1957) Technical change and the aggregate production function. Rev Econ Stat 39(3):312–320
Swan TW (1956) Economic growth and capital accumulation. Econ Rec 32(2):334–361
Tanzi V, Dawoodi H (1997) Corruption, public investment and growth. IMF, working paper 97/139
Uzawa H (1965) Optimal technical change in an aggregative model of economic growth. Int Econ Rev 6(1):18–31
Young A (1928) Increasing returns and economic progress. Econ J 38(152):527–542

Economic Impossibility

▶ Impracticability

Economic Integration

George Tridimas
Department of Accounting, Finance and Economics, Ulster University Business School, Belfast, Northern Ireland

Abstract

Economic integration is the establishment of a unified economic area where consumers and producers of different nations transact freely in a single market. Using the experience of the European Union, this essay offers a bird's–eye view of the trade–offs encountered when supranational structures pursuing collective objectives of integration may infringe on national sovereignty. The range of issues examined include: (A) Determination of policy with multiple veto players. (B) The advantage and disadvantages from centralising policy making. (C) The welfare effects of a customs union from changing the flows of trade and factors of production across different countries. (D) The costs and benefits from adopting a single currency and its consequences for budgetary policy.

Definition

Economic integration is the creation of a unified economic area where firms and consumers from different nations buy and sell goods and services in a single market and owners of capital and labor can deploy their resources in any economic activity anywhere in the area. Integration encompasses economic, political, and legal dimensions that overlap with each other. Our understanding and assessment of economic integration is inextricably linked to the experience gained from the establishment, geographical expansion, and extension of functions of the European Union (EU). Its development has been extensively researched in economics, political science, international relations, organizational sociology, and law. The present essay does not aim to survey any part of this enormous literature.[1] Rather, it offers some pointers on the issues regarding the following issues: setting up of supranational structures that pursue collective objectives but in so doing may encroach on national sovereignty, determination of policy with multiple veto players, centralization of policy making, market competition, trade effects of a customs union, factor mobility, and adoption of a single currency, monetary and budgetary policy.

European Economic Integration

The construction of the EU started after the end of Second World War and is ongoing. Table 1 presents a brief timeline of landmark events in the development of the EU. To complete the

[1]For textbook expositions, see amongst others Senior Nello (2011), Baldwin and Wplosz (2012) and Saurugger (2013). Detailed analysis of monetary integration can be found in Issing (2008) and De Grauwe (2012). The interested reader is also referred to the papers in the volume edited by Artis and Nixson (2007).

Economic Integration, Table 1 The evolution of the EU at a glance

1952	Belgium, France, Italy, Luxembourg, the Netherlands, and West Germany form the European Coal and Steel Community
1957	The six sign The Treaty of Rome, forming the European Economic Community (EEC) and the European Atomic Energy Community (Euratom)
1966	The Luxembourg Compromise is accepted, permitting member states to demand legislation to be adopted by unanimity when very important interests are at stake
1967	The three communities are united as "European Communities"
1973	Britain, Ireland, and Denmark join
1978	The European Monetary System and the European Currency Unit are founded
1981	Greece joins
1986	Spain and Portugal join
1986	The Single European Act is adopted, creating the Single Market
1990	Eastern Germany joins after the German reunification
1992	The Maastricht Treaty is signed, creating the European Union
1995	Austria, Sweden, and Finland join
1997	The Amsterdam Treaty is signed, amending the Maastricht Treaty
1998	The European Central Bank (ECB) is set up
1999	The *euro* currency is created by irrevocably locking the exchange rates of participating countries and monetary policy making is transferred to the ECB
2000	The Treaty of Nice Treaty is signed, amending earlier Treaties
2002	The *euro*" becomes the sole currency of 12 out of 15 members (Britain, Sweden, and Denmark retain their national currencies)
2004	Poland, Hungary, the Czech Republic, Slovakia, Slovenia, Latvia, Lithuania, Estonia, Malta, and (the Greek part of) Cyprus join
2004	The Treaty of Rome establishing a constitution for Europe is signed
2005	Dutch and French voters reject the 2004 Constitutional Treaty; EU leaders suspend its ratification
2007	Bulgaria and Romania join
2007	The Treaty of Lisbon is signed, amending the Treaty on European Union and the Treaty establishing the European Community
2009	The Treaty of Lisbon enters into force after its ratification is completed
2013	Croatia joins

economic union, a number of intermediate stages must be accomplished:

1. Duty-free access to each other nation's market, which requires the abolition of tariffs and non-tariff restrictions on trade between the member states.
2. Implementation of a common tariff on trade with non-member states to prevent cheaper imports entering the market through members states with lower external tariffs, known as a customs union.
3. Establishment of a free market in goods and services among the member states to ensure competition; this requires the harmonization of national laws and practices that regulate the market (including the relevant taxes), the abolition of anticompetitive structures, and the prohibition of state aid or other preferential treatment by member state governments to national firms.
4. Establishment of a free market in labor and capital, which is achieved by eliminating discriminatory treatment of workers on the basis of nationality, granting firms the right to establish in another member state and the removal of restrictions on capital flows.
5. Establishment of a monetary union where the member states adopt a single currency, so that prices and trading in the single market will not be affected by national currency fluctuations; in turn a single currency requires establishing a union-wide central bank to conduct monetary policy.

Each successive stage envelops its predecessor. Not all member states participate in the monetary union, with Denmark, Sweden, and the UK deciding to keep their national currencies, while countries that entered the union after the launch of the euro are expected to adopt the currency when their economies are ready to do so. A further step, one yet to be taken by the EU, is the fiscal union where taxes and transfers are decided centrally. It is however noted that contrary to the above checklist, the EU has adopted a protective agricultural policy. Each successive step implies a certain loss of national sovereignty, for example, a common tariff prevents a country to decide its own external trade policy or, with a common currency, a country can no longer implement an independent monetary policy.

Institutions of EU Governance

After the horrors of two world wars in a space of 20 years, the European economic integration was conceived as a form of international organization to bind together rival European powers, especially France and Germany, in order to prevent another war. That is, economic means were used to accomplish a political objective. What followed was a series of international treaties that over time established an intricate system of Europe-wide governance (control of decision making), changed the scope of national legislative powers, and created a body of legal acts and court decisions (known as the *acquis communautaire*) by which all member states abide, while on the economic side, it has shaped industry structures, labor market, trade, investment, and monetary flows.

Establishment of the EU by independent states meant the voluntary "pooling of sovereignty" to promote common interests and international public goods, like peace and prosperity, which are best pursued jointly rather than individually by each country. This also necessitated setting up bodies of collective decision making to pursue the common objectives and simultaneously governance mechanisms to check that the new bodies will act within the agreed limits without infringing on the rights of their creators. The main institutions established to drive and administer the process of integration are the following.

The European Council which consists of the leaders of the EU countries sets the EU's general political direction and priorities and deals with complex and sensitive issues that cannot be resolved at a lower level of intergovernmental cooperation. It meets four times a year and is chaired by the President of the European Council. The President of the European Commission and the EU's High Representative of the Union for Foreign Affairs and Security Policy also take part in the meetings. The decisions of the European Council are taken by unanimity or by qualified majority, depending on what the EU Treaty provides for. Though influential in setting the EU political agenda, it has no powers to pass laws.

The Council of the European Union (not to be confused with the previous European Council), also informally known as the EU Council, or Council of Ministers, which brings together national ministers from each EU country to pass EU laws coordinate the broad economic policies of EU member countries, sign agreements between the EU and other countries, approve the annual EU budget, develop the EU's foreign and defense policies, and coordinate cooperation between courts and police forces of member countries. As a general rule, the Council of the EU decides by qualified majority voting. From November 2014 a system known as "double majority voting" will be introduced. For a proposal to go through, it will need the support of 2 types of majority: a majority of countries (at least 15) and a majority of the total EU population (the countries in favor must represent at least 65% of the EU population). A blocking minority must include at least four Council members; if the latter fails, the qualified majority shall be deemed attained. When sensitive issues are decided, for example, security and external affairs and taxation, decisions have to be unanimous rendering veto powers to every single country. The Council of the EU represents the interests of the national governments of the member states.

The European Commission, which proposes policy measures, manages the day-to-day business of implementing EU policies and spending

EU funds and represents the EU internationally. It is an executive supranational body that represents the Community interests. It consists of 28 Commissioners, one from each country serving for a renewable term of 5 years. The President and members of the Commission are appointed by the European Council.

The European Parliament, which in an embryonic form, represents directly the interests of the peoples of Europe. Its role is to debate and pass laws in combination with the Council, debate and adopt the budget of the EU, and scrutinize other EU institutions. Its members are directly elected for terms of 5 years. Their number is 751 including its President; there is a minimum threshold of 6 members per country and a maximum of 96 (implying that smaller countries weigh higher in its composition). The members are grouped according to political affiliation and not by nationality. In comparison with national parliaments, it has significantly fewer legislative powers, but its powers have increased dramatically over time. It decides by simple majority.

The Court of Justice, whose task is to examine the legality of European Union measures and ensure the uniform interpretation and application of the EU law. It consists of one judge per EU country and eight "Advocates-General" whose job is to present opinions on the cases brought before the Court. Its members are appointed upon the common accord of the governments of the member states for renewable six-year terms. By far the largest part of the workload of the ECJ is to hear direct actions and give preliminary rulings. Direct actions concern violations of the EU law and include enforcement actions, where the Court declares whether or not a member state has infringed or complied with EU law; actions for judicial review, where the Court may annul an act of an EU institution for violating the EU law or for failing to make decisions required of them; and actions for damages, where the Court may determine the liability of an EU institution. In preliminary rulings, the Court on the request of a national court interprets a point of the EU law; this way it ensures the uniform interpretation of EU legislation.

The European Central Bank has been set up to manage the euro and maintain price stability in the EU. More specifically, its tasks are the determination and implementation of monetary policy by setting key interest rates for the 17 countries that currently use the euro as their currency (Austria, Belgium, Cyprus, Germany, Estonia, Greece, Spain, Finland, France, Ireland, Italy, Luxembourg, Malta, the Netherlands, Portugal, Slovenia, and Slovakia), the conduct of foreign exchange operations, the holding and management of the official foreign reserves of the euro area countries, and the promotion of the smooth operation of payment systems. The ECB is also responsible for framing and implementing the EU's economic and monetary policy. It is politically independent of the governments of the member states.

In comparing the role of the nation-states vis-à-vis that of the supranational bodies established to pursue and administer the objectives of European integration, two competing schools of thought appeared, namely, intergovernmentalism and supranationalism. Intergovernmentalism argues that the process of integration is controlled by the national governments of the member states which impose the policies that best suit their interests; hence, the institutions of international governance set up by the Treaties serve the purposes of their creators, usually the most powerful of the founding states, like France and Germany. On the contrary, supranationalism, or federalism, argues that economic integration, exchange, and cooperation across national borders generated a new transnational community whose interests are best served by setting institutional structures with some autonomous policy-making powers previously reserved for the nation-state.

Determination of Policy with Multiple Veto Players

The previous description of the governance organs makes clear that policy making in the EU is a complicated matter involving the strategic interaction of multiple players, where strategic means that the action of an actor takes into account the expected response of another actor whose interests are affected by such actions. The

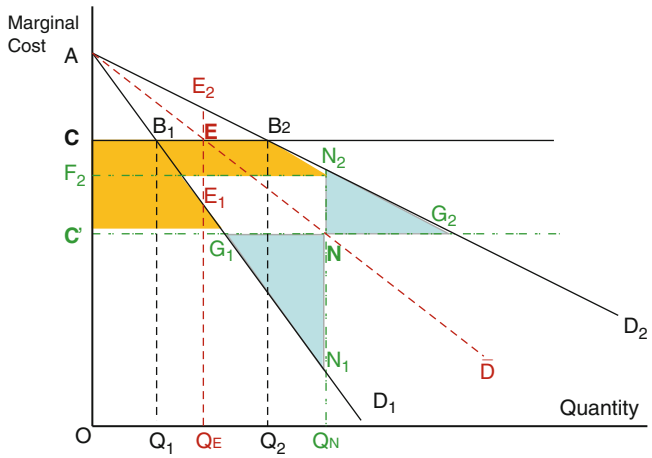
interests of the national and supranational actors may not necessarily coincide on all issues at hand, whereas only the Commission has agenda-setting powers to initiate legislation, and decisions are subject to the veto power of the Council, the Parliament (a process known as co-decision), and, as practice has shown, the Court.

The qualified majority voting rule used by the EU (in approximately 80% of all its decisions) in combination with its supranational character raises a host of important issues. The first regards efficiency in decision making, which relates to how easy it is for the EU as a collective group to take a decision. It refers to how likely is to find a majority given the specific voting rule and the distance between the policy preferences of the national and supranational actors. In general the answer to that depends on the required majority, the number of countries, and the weight (number of votes) of each country. The second issue relates to the distribution of power among member states is approximated by the weight awarded to each member state in the Council of Ministers. In the EU, the countries with small populations, like Luxembourg, have been given greater weights than the more populous countries, like Germany. This goes a long way to explain the observed pattern of EU spending in favor of less populous countries. The third issue regards legitimacy of the EU. A decision is accepted as legitimate when it is accepted that the decision maker has the right to take that decision. At one extreme, if the EU is considered as a union of states, as in a confederation, then legitimacy requires one vote per state. At the other extreme, if it is a union of peoples, then legitimacy requires equal power per citizen entitling more populous countries with increasing voting weight in the Council. Historically, the union-of-the peoples approach has been the EU norm since more populous states have a larger voting weight.

E

Centralization Versus Decentralization of Policy Making

Economic integration requires that some policies are decided by the supranational institutions, "the central authority," and a common policy applies to all member states. Examples include external tariff, market competition, environmental protection, arguably, monetary policy (if a single currency is adopted), financial regulation, and even some aspects of budgetary policy. This "harmonization" of policy opens up the debate of the merits of centralization versus decentralization. When the member states have different preferences for those policies but are forced to consume a greater (or smaller as the case may be) level of the service than it would have been optimal given their preferences when acting independently, a welfare loss results. This is depicted in Fig. 1 which shows the



Economic Integration, Fig. 1 Diversity of preferences, economies of scale, and welfare of centralization

demand for a public service by two countries *1* and *2* and D_1 and D_2, respectively, when the supply (marginal cost) of provision is C. Under decentralization the two countries act independently and consume Q_1 and Q_2 corresponding to the intersection of demand and supply. When they form a union, the central authority equates average demand to supply and provides the same quantity Q_E to both *1* and *2*. Since $Q_1 < Q_E < Q_2$, that is, country *1* overconsumes and country *2* underconsumes the public service, centralization generates the welfare losses measured, respectively, by the decrease in consumer surplus EB_1E_1 and EB_2E_2. Such losses depend on the diversity of preferences and the elasticity of demand (the distance between the demand curves and their slopes). Had the central authority the relevant information about the different preferences in the two countries, it would have been able to provide them with the individually optimal levels of the service; in the latter case, however, there would be no reason to form a union and centralize policy making.

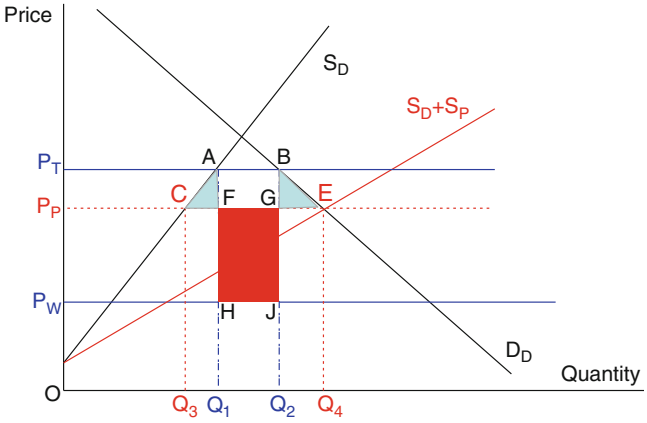
A second argument in favor of decentralization relates to the political benefits it offers. Specifically, citizens of independent states can choose their own government, so that they have the incentive and opportunity to make informed decisions about policies that affect them and they exercise more effective control over the discretionary powers of politicians. This way the political principal–agent problem is mitigated and accountability of politicians to voters is improved. In response to this issue, the EU has adopted the principles of *subsidiarity* and proportionality. Subsidiarity means that a policy is assigned to the supranational (=central) authorities when it cannot be achieved by the national authorities. According to proportionality, the content and form of the EU action shall not exceed what is necessary to achieve the objectives of the Treaties.

By contrast, the arguments in favor of centralization underline the benefits from economies of scale, correction of externalities, and elimination of inefficient noncooperative behavior. Centralization often involves significant economies of scale where increasing all inputs by the same proportion increases output more than proportionally and costs rise more slowly than production, resulting in important gains for the consumer. Graphically, the presence of economies of scale that are exploited under centralization yields a supply curve at C', lower than C. The new equilibrium obtained by the intersection of C' and is denoted by point N yields a larger equilibrium quantity for the two countries $Q_N > Q_E$. The gains from cost savings for countries 1 and 2 are shown, respectively, by the areas CB_1G_1C' and $CB_2N_2F_2$ and the corresponding losses from over- and underconsumption are represented by the triangles NN_1G_1 and NG_2N_2. Thus, the net welfare gains from centralization for 1 and 2 are given by the differences $CB_1G_1C'–NN_1G_1$ and $CB_2N_2F_2–N_2G_2N$.

Centralized decision making is also better equipped to address problems of positive or negative externalities (or spillovers), that is, situations where actions taken in one country (like burning fossil fuels or fishing) may increase or decrease the welfare of another country. However, the presence of externalities does not necessitate centralization, since such problems can be addressed by cooperation between the parties concerned. A third argument in favor of centralization is its ability to avoid noncooperative behavior by different countries. Countries competing against each other to attract business and mobile factors of production in their territories, which in turn increase the tax basis, may engage in games of competitive tax rate reductions (or other cost-cutting incentives, like relaxation of health and safety standards at the workplace) ultimately resulting in lower taxes and lower welfare in what is referred to as a "race to the bottom," that is, the lowest tax rate or protection for the workforce.

Trade and Growth Effects of Economic Integration

The economic rationale of integration among sovereign nations is that it promotes trade and growth and hence it increases welfare. In the short run, abolition of trade barriers allows economies to specialize according to their comparative



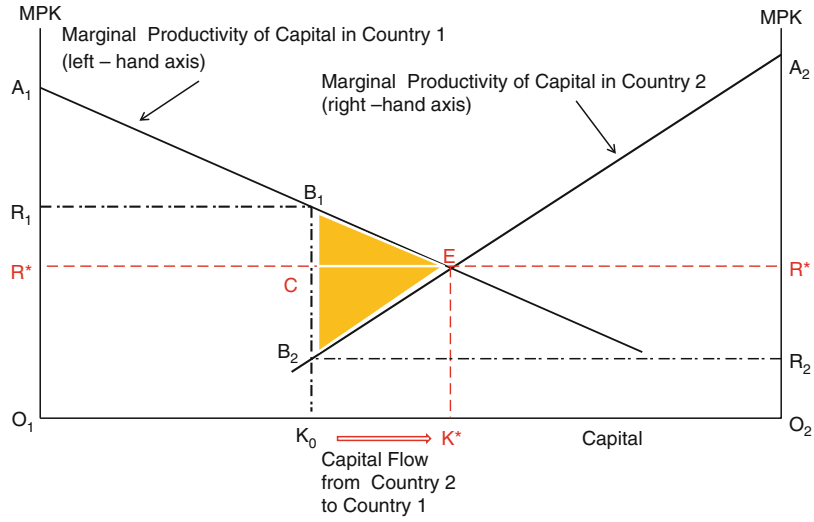
Economic Integration, Fig. 2 Trade and welfare effects of a customs union

Net welfare effect of preferential trade: ACF + BEG – HJGF

advantage which increases the volume of trade; however, against such benefits, one must set the costs from trade protection measures against non-members. These static effects of economic integration for an importing country are sketched in Fig. 2.

The lines D_D, S_D, and P_W denote, respectively, the domestic demand for the imported good, the domestic supply, and the world supply at the exogenously given P_W price. When the country applies a tariff to imports from all countries, the domestic price rises to P_T; domestic supply and demand are shown by OQ_1 and OQ_2, imports by Q_1Q_2, and import expenditure by Q_1HJQ_2. When a trade union is formed with a partner country that excludes other countries, domestic supply is represented by the line $S_D + S_P$ and equilibrium is obtained at E where D_D intersects $S_D + S_p$. Domestic demand rises to OQ_4, and domestic supply falls to OQ_3, implying the larger Q_3Q_4 volume of imports. Consumer welfare has increased by the sum ACF + BEG; this represents the gain from lowering the import price and is known as the trade creation effect. However, the Q_1Q_2 imports now enter the country at the P_P rather than P_W price paid to the partner country implying that the expenditure on Q_1Q_2 is HJGF. This is known as the trade diversion effect. Thence, the net welfare effect of forming the union is ACF + BEG – HJGF, which can be positive or negative. In other words, it is not a priori clear whether the preferential trade arrangement will benefit or harm the country members.

The effect of factor mobility is shown in Fig. 3 using the example of capital mobility. Assume again a two-country setting, where the capital endowment of countries 1 and 2, respectively, are O_1K_0 and K_0O_2 (giving a total capital stock of O_1O_2 as shown on the horizontal axis). Lines MPK_1 and MPK_2 show the marginal productivity of capital in the two nations (or, equivalently, the national demand for capital functions), and R_1 and R_2 show the return on capital in the two countries before capital market integration, with $R_1 > R_2$. The areas defined by the MPK curves, the axis and the individual capital endowments, $O_1A_1B_1K_0$ and $O_2A_2B_2K_0$, show the outputs of countries 1 and 2, respectively. $A_1B_1R_1$ and $O_1R_1B_1K_0$ represent the sizes of labor and capital income in country 1, and $A_2B_2R_2$ and $O_2R_2B_2K_0$ show the corresponding magnitudes in country 2. When capital markets are integrated, capital flows freely from the low-return country 2 to the high-return country 1 until returns are equalized at R* at the intersection of MPK_1 and MPK_2; country 1 ends up using domestically capital O_1K^* (although it owns OK_0), total output equal to $O_1A_1EK^*$, and labor income A_1ER^* (higher than before by $R_1B_1ER^*$). In country 2, output falls to $O_2A_2EK^*$, labor income falls to A_2ER^*, and capital income rises to $O_2R^*CK_0$ (upon adding the capital earnings from capital owned by country 2 but used in 1). Thus, the net effect of integration is again



Economic Integration, Fig. 3 The effects of factor mobility

Net welfare gains from capital market integration: $B_1EC + B_2EC$

represented by B_1EB_2, which is the sum of the extra labor income in country 1, B_1EC, and the extra capital income in country 2, B_2EC. Clearly, although the overall integration has a positive effect on welfare, it creates winners and losers in each country, who may resist it with various degree of success depending on their political influence. Similar conclusions about increased output and its differential distribution are derived when we consider the labor market and allow immigration from a low-wage to a high-wage country.

Over the long run, integration implies that firms have access to a larger union-wide market and that they face stiffer competition than when they were confined to their national market only. As a result of more competition, costs and profit margins are squeezed and prices fall. Firms that survive become bigger and better able to exploit economies of scale driving prices further down to the benefit of consumer. The benefits of the restructuring process, however, are at risk from two sources. National governments bowing to political pressures may subsidize and otherwise assist failing firms, delaying the efficient restructuring of the economy. Secondly, the emergence of fewer but bigger firms may lead to price collusion, negating the benefits of lower prices for consumers. It is for these reasons that the EU has introduced strict rules forbidding state aid and enforcing competition. Since integration encourages a more efficient allocation of resources, human and nonhuman, labor and capital are allocated more efficiently across different countries, boosting the rate of economic growth.

Monetary Union

Monetary union or monetary integration is an arrangement between participating countries where the exchange rates are permanently and irrevocably fixed, so that a single currency can be used by all members. Several benefits are associated with the adoption of a single currency:

(a) The elimination of exchange rate fluctuations and the ensuing uncertainty which discourages trade and investment.
(b) Reduction in transaction costs relating to conversion fees and commission charges incurred when exchanging different currencies.
(c) Seignorage gains from establishing the single currency as an international reserve currency; these arise from the willingness of the rest of the world to hold the single currency as an asset which then allows the monetary union to import more than it exports.

(d) Reduction in the opportunity cost of keeping foreign reserves, since the union needs fewer reserves to manage its currency than the sum of reserves needed by each member state acting independently.
(e) Greater effectiveness in pursuing aggregate stabilization policy. While many small open economies acting on their own cannot implement successful short-run aggregate demand management policies because of their dependence on international trade, a union can succeed by coordinating policy among member states.
(f) For countries characterized by high inflation before the formation of the monetary union (like Italy and Greece), adopting a single currency managed by a central bank which is committed to price stability introduces a credible anti-inflationary assurance (or so the argument ran before the debt crisis of 2010).

However, such benefits may be accompanied by severe costs: giving up monetary policy independence and adopting the single currency imply that a country can no longer use the exchange rate to counteract demand and/or supply shocks. When domestic prices and wages are slow to adjust to such shocks, the exchange rate can be used to adjust domestic demand and supply to reestablish macroeconomic equilibrium quickly. With a floating exchange rate system, the exchange rate will adjust to maintain balance of payments equilibrium; monetary policy becomes effective to affect the domestic economy, but fiscal policy becomes ineffective (because capital will move in and out of the country responding to domestic and foreign interest rate differentials). On the other hand, under a fixed exchange rate regime, the authorities must respond to a surplus by reflating and to a deficit by deflating; thus, monetary policy becomes ineffective, while fiscal policy is now effective. These considerations point to the "impossible trinity" of having simultaneously a fixed exchange rate, independent monetary policy, and perfect capital mobility.

Whether or not a group of countries will benefit from forming a monetary union has been examined by the "optimum currency area" (OCA) literature. This starts from the observation that the larger the area using the same currency, the larger its benefits; but as the area grows larger, it includes more diverse countries, which increases the costs of using the currency. Specifically, in the face of an adverse external shock that affects different countries differently (small economic losses for some but large for others), a single central bank cannot differentiate its policy responses according to the needs of the different countries. Asymmetric costs may deter the formation of a currency union. A group of countries can form a successful OCA when the following conditions are satisfied: (a) The labor force is mobile across the different countries. (b) The countries have diversified economies and produce and export similar goods. (c) The countries are open to international trade with each other. (d) They have adopted a mechanism of fiscal transfers that compensate each other for adverse economic shocks. (e) They share common preferences on how to respond to an external shock. (f) Perhaps more importantly, since none of the previous criteria may be fully satisfied, the countries have a sense of solidarity that their fates bound them together and so accept the costs of asymmetric shocks.

When countries lose the exchange rate as a policy instrument, they can use fiscal means to counteract any adverse shocks to aggregate demand, especially if fiscal transfers are not feasible (see (d) above). Despite the efficacy of discretionary fiscal policy in a system of fixed exchange rates, its use by member states of an economic union with a single currency is controversial. Acting independently, a member state running persistent large budget deficits adding to national debt may lead to severe difficulties. To bail out the highly indebted country, the central bank may increase the money supply, bringing inflation and depreciation of the single currency. Alternatively, debt accumulation may lead to higher interest rates for all country members crowding out private investment. If capital markets are efficient and assess the default risks of different governments, they will demand significantly higher interest on the debt of the profligate

government without a general interest rate increase. But if high indebtedness leads to fears about the financial stability of other countries, the commitment of the central bank to low inflation may no longer be believed, implying that the policy of no bailouts for fiscal profligacy lacks credibility. Accordingly, the risk of default is lower than otherwise. This generates a moral hazard problem where a member has an incentive for spending profligacy. Note that there may even be an adverse selection problem where only the worst offenders (those who run the biggest budgetary deficits, like Italy and Greece) are interested in joining the monetary union.

Cross-References

- ▶ Court of Justice of the European Union
- ▶ European Nationality
- ▶ Efficiency
- ▶ Externalities
- ▶ Fiscal Federalism
- ▶ Power Indices
- ▶ Simple Majority

References

Artis M, Nixson F (2007) Economics of the European Union, 4th edn. Oxford University Press, Oxford

Baldwin R, Wyplosz C (2012) The economics of European integration, 4th edn. McGraw Hill, Maidenhead

De Grauwe P (2012) The economics of Monetary union, 7th edn. Oxford University Press, Oxford

For annual scholarly updates on EU developments the reader is referred to the Supplement of the Journal of Common Market Studies, an academic publication dedicated to EU issues, The interested reader may also consult the EU website: http://europa.eu/index_en.htm (in English)

Issing O (2008) The birth of the Euro. Cambridge University Press, Cambridge

Saurugger S (2013) Theoretical approaches to European integration. Palgrave Macmillan, Basingstoke

Senior Nello S (2011) The European Union: economics, policies and history, 3rd edn. McGraw Hill, Maidenhead

The stock of scholarly work on the economic, political and legal aspects of European integration is enormous and, in view of the fast pace of the changes recent change, expanding rapidly. The following list is only a small sample of some of the most popular texts

Economic Performance

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

Abstract

Economic performance indicates the way in which a country or a firm functions, that is the efficiency with which they achieve their intended objectives. Alternative approaches to measuring both macroeconomic and microeconomic performance exist. Gross Domestic Product, among others, is one of the most important indicator to know how well an economy is performing and, given the existence of international standards for its calculation, it is also widely used across the world for international comparison. Efficiency and productivity, as well as long-term growth, are traditional indicators of firm performance. It is important to note, however, that all indicators are imperfect and, overall, it is important for economists and policy makers to look beyond the headline statistics to give a better overall picture of economic welfare.

Definition

Economic performance indicates the way in which an economy functions. In general, it signals the efficiency with which a country or a firm achieves its intended purposes.

What we intend to measure

> . . .Every day, in every industrialized country of the world, journalists and politicians give out a conscious and unconscious message. It is that better economic performance means more happiness for a nation. This idea is rarely questioned.. . . (Oswald 1997)

A prerequisite for examining alternative approaches to measuring economic performance

is an agreement on precisely what we intend to measure. This objective is closely related to the goals that societies and, hence, policy makers want to pursue. However, there is little doubt that economic policy makers need regular, timely, and accurate indicators of economic performance which, on the other hand, can relate to both countries (macroeconomic performance) and firms (microeconomic performance).

Measuring Macroeconomic Performance

The performance of an economy is usually assessed in terms of the achievement of economic objectives. These objectives can be long run, such as sustainable growth and development or short term, such as the stabilization of the economy in response to sudden and unpredictable economic shocks.

In general, to know how well an economy is performing, economists employ a wide range of economic indicators. Economic indicators measure macroeconomic variables that directly or indirectly enable economists to judge whether economic performance improves or deteriorates. Tracking these indicators is especially valuable to policy makers, both in terms of assessing whether to intervene and whether the intervention has worked or not.

Traditionally, the key measures of macroeconomic performance include:

1. Levels of real national income, spending, and output, three key variables that indicate whether an economy is growing or in recession. Like many other indicators, they are usually measured in per capita (per head) terms.
2. Economic growth, real gross domestic product (GDP) growth, GDP per capita.
3. GDP per hours worked as a measure of economic productivity, labor productivity.
4. Price levels and inflation.
5. Employment rate for the 15–64 age group and patterns of employment.
6. Unemployment levels and types.
7. Current account – satisfactory current account (e.g., low deficit).
8. Balance of payments.
9. Final consumption expenditure per capita, including government consumption.
10. Income inequality.
11. Investment levels and the relationship between capital investment and national output.
12. Levels of savings and savings ratios.
13. Competitiveness of exports.
14. Trade deficits and surpluses with specific countries or the rest of the world.
15. Debt levels with other countries.
16. The proportion of debt to national income.
17. The terms of trade of a country.
18. The purchasing power of a country's currency.
19. Wider measures of human development, including literacy rates and health-care provision. Such measures are included in the Human Development Index (HDI).
20. Measures of human poverty, including the Human Poverty Index (HPI).

Gross Domestic Product (GDP)

GDP is one of the most important indicators of economic performance. Since there are international standards for its calculation, it is also widely used across the world for international comparisons.

As GDP is a measure of a country's overall production for any given year, it is a reliable, albeit still imperfect, gauge of a country's economic performance. This is the justification for the great attention which both the general public and policy makers pay in all advanced economies to the regularly published GDP figures.

GDP, however, has several limitations.

Firstly, GDP doesn't take into account income distribution, since it could primarily benefit the top income strata of the population.

Secondly, it is characterized by well-known deficiencies related to the measurement of economic activities. Indeed, various nonmarket outputs, such as household activities and services provided free of charge, are systematically overlooked. The underground economy is difficult to capture, particularly certain criminal activities, although several attempts have been made to

harmonize the coverage of the underground economy at EU level in order to obtain comparable GDP measures. Some elements of GDP are fragile estimates, particularly those of the volume of publicly provided services and of the quality incorporated into products. Finally, some expenditures are unequivocally counted as positive contributors to economic performance, while the negative externalities associated with them – such as environmental damage – are neglected. As a result, GDP understates output.

Finally, even if GDP is a measure of market production, it has often been treated as if it were a measure of economic well-being. The measurement of GDP, however, does not address all aspects which are relevant for the material well-being of an economy; thus, it is not always a reliable guide to living standards. While the general public and many policymakers regard GDP as a measure of material well-being, this interpretation ignores the fact that production is not the ultimate goal of a society. In this context, production-based measures need to be complemented by a broader set of indicators if the aim is to assess well-being. These additional indicators would include social investment like infrastructure, education, access to health care, housing, as well as quality of life like material wealth, mental state, stress, perception of crime, environment, etc. Conflating GDP and well-being can lead to misleading indications about how well-off people are and entail the wrong policy decisions.

Making GDP a Better Measure of Economic Performance

Economic policy makers unquestionably need an economic performance indicator for short-term decision-making. Macroeconomic policy frequently operates with a time horizon of 1–2 years and, from this perspective GDP, as an indicator of current value added, is arguably the most informative gauge of economic performance. However, even in this area of economic policy, it is necessary to go "beyond GDP" by analyzing data on unemployment, inflation, short-term business activity, and consumer or business sentiment.

Although the usefulness of GDP is limited from a medium-term perspective, it still remains a viable indicator of medium-term performance. Thus, in conceptual terms, GDP remains the cornerstone of economic performance assessments. Nevertheless, it should be improved in various directions.

The most important starting points for improvements are (i) improving the measurement of service output in general and of government services in particular and (ii) making progress in measuring quality improvements.

Employment Rate

Unemployment rate is another important indicator of economic performance. However, it is heavily influenced by country-specific legislation and programs to combat joblessness. Moreover, whenever unemployment is too high and long-lasting, workers might quit the labor market, making intercountry comparisons particularly unreliable.

A more direct indicator could be the probability of being employed at working age. The employment rate in the population aged 15–64 years is often used. This basic indicator has already gained widespread acceptance in labor economics and statistics.

While such an indicator admittedly does not tell us anything about job quality or whether jobs match people's expectations, it nevertheless means a lot when looking for a job or being exhausted by long periods of job search. It is also a sustainability indicator as it is an important parameter for the long-term future of retirement plans and public finances.

All Statistics Are Limited

It is important to note that all statistics have some limitation.

Real GDP will always be useful for showing the stage in the economic cycle. It is of some use in indicating living standards. But, it is far from the ultimate guide. Even employment rates can be partially misleading. For example, is the employment temporary or permanent? Employment figures have been better than expected, but there has been a rapid rise in labor market insecurity as well.

Therefore, there is always a need to look at several statistics at the same time, and, overall, it is important for economists to look beyond the headline statistics to give a better overall picture of economic welfare.

Measuring Microeconomic Performance

Microeconomic performance signals a firm's success in areas related to its assets, liabilities, and overall market strengths. To know how well a firm is performing, economists employ some economic indicators that directly or indirectly enable economists to judge whether economic performance has improved or deteriorated. These indicators usually include:

1. Efficiency and productivity
2. Long-term growth

Efficiency indicates that production proceeds at the lowest possible per-unit cost. Productivity is an average measure of the efficiency of production. It can be expressed as the ratio of output to inputs used in the production process, i.e., output per unit of input.

Note that productivity growth, in particular, is seen as the key economic indicator of innovation and it is a crucial factor in production performance of both firms and nations. Increasing national productivity can raise living standards because more real income improves people's ability to purchase goods and services, enjoy leisure, improve housing and education, and contribute to social and environmental programs. Productivity growth also helps businesses to be more profitable.

Firms' performance, particularly measured by efficiency and productivity, is a long-standing topic in economic studies (Coelli et al. 2004; OECD 2001).

Research developments in this direction are mostly related to methodological advancement. The empirical literature on firm efficiency has been dominated by a nonparametric approach – the data envelopment analysis (DEA) (Charnes et al. 1994; Seiford 1996). The main advantages of DEA compared to the standard econometric technique are that: (1) it does not require any form of functional specification; (2) it is able to handle multiple inputs and outputs readily in any (in)efficiency theoretical paradigm. Based on the input and output data from DEA, a Malmquist Index can be constructed to measure productivity change. The criticism of the DEA method is related to its potential statistical shortcomings. A further development of this method is to use the bootstrap approach to obtain statistical properties.

Another well-developed method is the stochastic frontier approach (a parametric approach) (Aigner et al. 1977). Its principal advantage lies in the decomposition of deviations from the efficiency levels between noise (stochastic error) and pure efficiency; however, it faces the challenge of determining the appropriate functional forms. Recently, a semi-parametric method which combines nonparametric and parametric approaches has been applied (Bernini et al. 2004).

In addition to efficiency and productivity, firms' long-term growth – that is, sales, turnover, or profit growth – is also used to measure firm performance. Based on the production function (i.e., input–output), significant input factors are identified to explain the growth. This line of research departs from the SCP paradigm but does not seek explanations of firm growth from market structure or conduct.

References

Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production functions. J Econ 6:21–37

Bernini C, Freo M, Gardini A (2004) Quantile estimation of frontier production function. Empir Econ 29(2):373–381

Charnes A, Cooper WW, Lewin AY (1994) Data envelopment analysis. Kluwer, Boston/Dordrecht/London

Coelli TJ, Rao DSP, O'Donnell CJ, Battese GE (2004) An introduction to efficiency and productivity analysis. Springer, New York

OECD (2001) Measuring productivity, OECD manual. OECD Publishing, Paris

Oswald AJ (1997) Happiness and economic performance. Econ J 107(445):1815–1831

Seiford LM (1996) Data envelopment analysis: the evolution of state of the art (1978–1995). J Prod Anal 7:99

Economics Imperialism in Law and Economics

Magdalena Małecka
TINT – Centre for Philosophy of Social Science, University of Helsinki, Helsinki, Finland

Definition

Economics imperialism is a type of an interdisciplinary relation between the scientific discipline of economics and other disciplines of the social sciences. The disciplines that are often claimed to be "imperialized" by economics are sociology (see Granovetter and Swedberg 2001; Swedberg 1990), political science (see Hodgson 1994; Sigelman and Goldfarb 2012; Kuorikoski and Lehtinen 2010), anthropology (see Marchionatti and Cedrini 2016), geography (see Mäki and Marchionni 2010), and law (see Medema 2018; Davies 2010; Fink 2003). Hence, the law and economics movement is seen by some as one of the manifestations of the imperialism of economics in the social sciences (in this case – in legal scholarship). Below, I review the general debate on economics imperialism, and then I discuss the attempts to analyze law and economics as a case of economics imperialism.

Introduction

Recently, in the philosophy of science literature there is a discussion on scientific imperialism, of which economics imperialism is just an instance. This debate has revolved around the question of the permissibility of the application of scientific theories and methods outside the discipline in which they were initially introduced. Philosophers of science have attempted to clarify what it means for a theory or a discipline to be applied outside its own field or domain, and whether such an application can be understood as imperialistic (Dupré 1994, 2001; Mäki 2013; Clarke and Walsh 2009; Kidd 2013; Clarke and Walsh 2013; contributions to the volume edited by Mäki et al. 2018). It is debated whether scientific imperialism is a neutral or normatively loaded term, how to establish criteria for assessment of imperialistic practices in science, how to categorize different instances of scientific imperialism, as well as whether the political metaphor of imperialism is useful at all in order to account for relationships between scientific disciplines. Mäki et al. (2018) argue that scientific imperialism challenges two central tenets of current scientific practice, as it calls into question two widely spread ideas: (1) that the broader the scope of a scientific theory, the better; (2) that interdisciplinary exchanges bring indisputable benefits (p. 1).

The term "economics imperialism" has been used in the pejorative sense by the critics of the idea of applying economics outside its domain (Fine and Milonakis 2009; Davis 2015; Hodgson 1994; Nik-Khah and Van Horn 2012) and in the approbatory sense by some economists (most notably by Becker 1971; Hirshleifer 1985; Stigler 1984; Lazear 2000; Radnitzky and Bernholz 1987; Tullock 1972) who have advocated the idea of expanding economic theories and methods to the topics studied by the less advanced, in their opinion, social sciences. "Economics imperialists" have argued that such an expansion is justified by economics' "growing abstractness" and "generality" that economics achieves thanks to "the machine of maximizing behavior" (Stigler 1984); by the universal applicability of economics' analytical categories (Hirshleifer 1985); and by economics' intellectual rigor (Lazear 2000; Posner 1989). Moreover, Mäki (2009) emphasizes that the expansion of economics has been often justified in terms of the ideal of unification – economics is supposed to provide, in the opinion of "imperialists," a unifying theory for the social sciences. Fine and Milonakis (2009) claim that the theoretical development in economics that made the expansion of economics possible was the marginalist revolution and its basic concepts of equilibrium (see: ▶ "Equilibrium Theory"), rationality (see: ▶ "Rationality"), scarcity – "universal in content and application" (p. 8).

Criticism of Economics Imperialism

However, critics of economics imperialism, apart from questioning the alleged advancement of

economics as a social science, often point out the importance of non-epistemic arguments and non-academic standing of economics for the successes of its imperialistic scientific practices. For instance, Nik-Khah and Van Horn (2012) claim that economics imperialists of the Chicago School wanted in the first place to influence policy approach by proposing their view on what constitutes the economic analysis of state policy. "Stigler's program to study 'governmental control' constituted a new and distinct approach, rather than mere application of a core Chicago approach to a new domain" (p. 217). Amadae (2018) stresses the importance of politically motivated expansion of the rational choice theory – its attractiveness for nuclear deterrence made it prestigious in nonacademic contexts and in this way has strengthened its standing within academia as well. In this context the affinities of the Chicago School analysis with neoliberalism are often emphasized. Nik-Khah and Van Horn (2012) claim that economics imperialism cannot be really understood if one does not take into account the Chicago School economists' involvement into the revival of the liberal project, known as neoliberalism – redefinition of the social and the political sphere through the concepts of neoclassical economics – and its impact on the real-world policy design. Therefore, partly in response to this "performative" aims of the Chicago School (Davis 2015), or "pedagogical" effects of the rational choice theory (Amadae 2018), some critics of economics imperialism have argued that the conceptualization of social relationships through the categories of neoclassical economics, or of rational choice, is inherently inadequate (Fine and Milonakis 2009) and can lead to ethically objectionable large-scale sociocultural consequences (Marino 2018) by influencing agents' self-understanding (Clarke and Walsh 2009).

Is All Economics Imperialistic?

It should be noticed, however, that upon inspection we find out that economics imperialism, as each instance of scientific imperialism, is in fact a relationship between smaller epistemic units than scientific disciplines (Davis 2015; Małecka and Lepenies 2018; Chassonnery-Zaïgouche 2018). The analysis of cases of economics imperialism makes it clear that it is never the case that the *whole* discipline of economics is "imperializing," e.g., the *whole* sociology. Thus, as Davis (2012) argues, "economics is not all economics" (p. 210). It is usually a particular research program that develops within the discipline of economics and which is being transferred outside, as well as within, economics' institutional borders.

In contemporary discussions, it is often claimed that only neoclassical economics, and more specifically the Chicago School and its price theory, are extended to other domains and scientific fields (Davis 2015; Nik-Khah and Van Horn 2012). However, Amadae (2018) argues that it is in fact not the neoclassical economics with its formalization of diminishing marginal utility, what constitutes economics imperialism, but rather the game theory with its view on agency as strategic competition between actors who satisfy their preferences. Guilhot and Marciano (2018) complicate the picture by showing that what is now called the economics imperialism in political science was the application of the rational choice theory to the questions of the political decision-making: it was possible rather due to the rising importance of the decision theory across the social and behavioral sciences after the Second World War. Furthermore, few researchers emphasize that some of the economic approaches, or research programs that we consider imperialistic today, were in fact quite marginal within economics, back in the days. Becker's (see ▶ "Becker, Gary S.") application of rational choice and price theory to the analysis of social phenomena is one of the examples (Chassonnery-Zaïgouche 2018; Vromen 2009). Chassonnery-Zaïgouche (2018) analyzes economic studies of discrimination and claims that Becker influenced, imperialized, with his work mainly the discipline of economics and has had, in fact, marginal impact on other social sciences studying discrimination.

Despite the seemingly intuitive appeal of the term "economics imperialism," it is not clear what does it mean that one research program is imperializing another, or other. Economics imperialism has been defined, for instance, as

"colonisation of the subject matter of other social sciences by economics" (Fine and Milonakis 2009), as "a form of economics expansionism where the new types of explanandum phenomena are located in territories that are occupied by disciplines other than economics" (Mäki 2009), or as "the attempt to extend the core ideas of neoclassical economics to cover social science as a whole" (Hodgson 1994). Most of the definitions account for the metaphorical use of the notion of imperialism as we know it in political context. There is a discussion, however, to what extent the metaphor of imperialism should be taken seriously when talking about scientific practices. The notion of imperialism is normatively loaded and most scholars believe that one cannot ignore this normative dimension in the case of economics imperialism (Mäki 2013, who defines scientific, and economics, imperialism neutrally, is an exception here). Małecka and Lepenies (2018) provide the definition of scientific imperialism, that applies to economics imperialism, in which they account for the widely shared believe that there is something normatively problematic about economics imperialism. They stress the importance of epistemic and non-epistemic factors for being able to define and identify economics imperialism. Scientific imperialism is an activity that is related both to a certain view on the progressive character of a novel application of an existing research approach (the epistemic aspect) and to a power to favor this approach at the expense of other approaches in terms of academic and non-academic prestige, or/and resources (the institutional aspect).

Law and Economics as an Instance of Economics Imperialism

Law and economics is often seen as a result of the expansion of the Chicago School approach to law (Mercuro and Medema 2006; Medema 2015, 2018). Sometimes it is understood, more broadly, as an application of orthodox, or neoclassical economics (Jackson 1984; Davies 2010), or of game theory (Pearson 1997) to law. Mercuro and Medema (2006) even argue that law and economics "has been the most successful of economists' imperialistic forays into other disciplines" (p. 100).

The application of the price theory to the analysis of law is grounded on the premise that individuals are rational maximizers of their utility, that they respond to price incentives also in non-economic settings and that law can be treated as an incentive. Medema (2015) stresses the importance of Gary Becker's (see ▶ "Becker, Gary S.") works that advanced the analysis of all social phenomena with the tools of price theory (and econometrics) for the development of economic analysis of law. Medema (2015) differentiates this "new" law and economics, inspired by the work of Becker, as well as of Richard Posner (see ▶ "Posner, Richard"), from the "old" one, initiated by Aaron Director (see ▶ "Director, Aaron"), whose aim was to simply analyze the impact of legal rules on economic performance. According to Medema only the "new" law and economics has features of scientific (economics) imperialism (compare also Epstein 1997 on periodization of law and economics and Harnay and Marciano (2009) on the difference between law and economics and economic analysis on law).

Posner (1989) justifies the application of neoclassical economics to law by the rigor that this type of economic analysis allegedly offers, as well as by the possibility of funding the truly "scientific" analysis of law in this way. Cooter (1981) points out that the reason for the economics imperialism in law is the "discovery" by economists a niche in legal scholarship – a lack of quantitative reasoning. Cooter (1995) also emphasizes the importance of an attempt for unification for applying economics to law.

The historical case study made by Medema (2018) on the ad hoc Joint Committee of the American Economic Association and the Association of American Law Schools established in 1966 to explore the prospects for interactions between lawyers and economists challenges the commonly spread narrative about neoclassical economists conquering a new field. The study demonstrates that lawyers played a crucial role at the early stage of bringing economics to their studies and in replacing "the traditional methods of legal analysis" (Medema 2018, p. 110). However, later on

the application of neoclassical economics to law had been opposed by some lawyers as being too abstract an analysis, untested, irrelevant to the courtroom (Cooter 1981). Law and economics as a scientific project of explaining legal phenomena has been also scrutinized and criticized. For instance, Jackson (1984) criticized law and economics' scientism and its technocratic attitude that made it so dominant in the legal scholarship. For other critics law and economics is mainly a manifestation of "the neo-liberal project of applying neo-classical economics to state sovereignty" (Davies 2010, p. 64) that has been advanced by non-epistemic arguments and normative views on policy as complying with the goals of economic efficiency (Davies 2010; Fink 2003).

Cross-References

- ▶ Becker, Gary S.
- ▶ Equilibrium Theory
- ▶ Posner, Richard
- ▶ Rationality

References

Amadae SM (2018) Economics imperialism reconsidered. In: Scientific imperialism: exploring the boundaries of interdisciplinarity. Routledge, p 4

Becker GS (1971) Economic theory. Alfred A. Knopt, New York

Chassonnery-Zaïgouche C (2018) Crossing boundaries, displacing previous knowledge and claiming superiority. In: Scientific imperialism: exploring the boundaries of interdisciplinarity. Routledge, p 31

Clarke S, Walsh A (2009) Scientific imperialism and the proper relations between the sciences. Int Stud Philos Sci 23(2):195–207

Clarke S, Walsh A (2013) Imperialism, progress, developmental teleology, and interdisciplinary unification. Inter Stud Philos Sci 27(3):341–351

Cooter RD (1981) Law and the imperialism of economics: an introduction to the economic analysis of law and a review of the major books. UCLA Law Rev 29:1260

Cooter R (1995) Law and unified social theory. J Law Soc 22(1):50–67

Davies W (2010) Economics and the 'nonsense'of law: the case of the Chicago antitrust revolution. Econ Soc 39(1):64–83

Davis JB (2012) Mäki on economics imperialism. In: Economics for real: Uskali Mäki and the place of truth in economics. Routledge, pp 203–219

Davis JB (2015) Economics imperialism versus multidisciplinarity

Dupré J (1994) Against scientific imperialism. In PSA: proceedings of the Biennial meeting of the philosophy of science association, Philosophy of Science Association, vol 1994, no 2, pp 374–381

Dupré J (2001) Human nature and the limits of science. Clarendon Press, Oxford

Fine B, Milonakis D (2009) From economics imperialism to freakonomics: the shifting boundaries between economics and other social sciences. Routledge, London/New York

Fink EM (2003) Post-realism, or the jurisprudential logic of late capitalism: a socio-legal analysis of the rise and diffusion of law and economics. Hastings LJ, 55, 931

Granovetter M, Swedberg R (2001) The sociology of economic life, 2nd edn. Westview Press, Boulder

Guilhot N, Marciano A (2018) Rational choice as neo-decisionism. In: Scientific imperialism: exploring the boundaries of interdisciplinarity. Routledge

Harnay S, Marciano A (2009) Posner, economics and the law: from "law and economics" to an economic analysis of law. J Hist Econ Thought 31(2):215–232

Hirshleifer J (1985) The expanding domain of economics. Am Econ Rev 75(6):53–68

Hodgson GM (1994) Some remarks on 'economic imperialism'and international political economy. Rev Int Polit Econ 1(1):21–28

Jackson N (1984) The economic explanation of legal phenomena. Int Rev Law Econ 4(2):163–183

Kidd JI (2013) Historical contingency and the impact of scientific imperialism. Int Stud Philos Sci 27(3):315–324

Kuorikoski J, Lehtinen A (2010) Economics imperialism and solution concepts in political science. Philos Soc Sci 40(3):347–374

Lazear EP (2000) Economic imperialism. Q J Econ 115(1): 99–146

Mäki U (2009) Economics imperialism: concept and constraints. Philos Soc Sci 39(3):351–380

Mäki U (2013) Scientific imperialism: difficulties in definition, identification, and assessment. Int Stud Philos Sci 27(3):325–339

Mäki U, Marchionni C (2010) Is geographical economics imperializing economic geography? J Econ Geogr 11(4):645–665

Mäki U, Walsh A, Pinto MF (eds) (2018) Scientific imperialism: exploring the boundaries of interdisciplinarity. Routledge, Oxon

Małecka M, Lepenies R (2018) Is the behavioral approach a form of scientific imperialism? An analysis of law and policy. In: Mäki U, Walsh A, Pinto MF (eds) Scientific imperialism: exploring the boundaries of interdisciplinarity. Routledge, Abingdon

Marchionatti R, Cedrini M (2016) Economics as social science: economics imperialism and the challenge of interdisciplinarity. Taylor & Francis, Basingstoke

Marino P (2018) Ethical implications of scientific imperialism. Two examples from economics. In: Scientific imperialism: exploring the boundaries of interdisciplinarity. Routledge

Medema SG (2015) From dismal to dominance? Law and economics: philosophical issues and fundamental questions 22: 69
Medema SG (2018) Scientific imperialism or merely boundary crossing? In: Scientific imperialism: exploring the boundaries of interdisciplinarity. Routledge
Mercuro N, Medema SG (2006) Economics and the law: from posner to postmodernism and beyond, 2nd edn. Princeton University Press
Nik-Khah E, Van Horn R (2012) Inland empire: economics imperialism as an imperative of Chicago neoliberalism. J Econ Methodol 19(3):259–282
Pearson H (1997) Origins of law and economics: the economists' new science of law, 1830–1930. Cambridge University Press, New York
Posner R (1989) Foreword. In: Faure M, van den Bergh R (eds) Essays in law and economics: corporations, accident prevention and compensation for losses. Maklu, Antwerpen
Radnitzky G, Bernholz P (eds) (1987) Economic imperialism: the economic approach applied outside the field of economics. Paragon House Publishers, New York
Sigelman L, Goldfarb R (2012) The influence of economics on political science: by what pathway? J Econ Methodol 19(1):1–19
Stigler GJ (1984) Economics: the imperial science? Scand J Econ 86(3):301–313
Swedberg R (1990) Economics and sociology: redefining their boundaries: conversations with economists and sociologists. Princeton University Press, Princeton
Tullock G (1972) Economic imperialism. In: Theory of public choice. The University of Michigan Press, pp 317–29
Vromen JJ (2009) The booming economics-made-fun genre: more than having fun, but less than economics imperialism. Erasmus J Philos Econ 2(1):70–99

Ecosystem Services

Sébastien Roussel
Department of Economics, CEE-M, Université Montpellier, CNRS, INRA, SupAgro, Université Paul Valéry Montpellier 3, Montpellier, France

Definition

Ecosystem services are services provided by the ecosystems and contributing to human well-being. These services play a crucial role in signaling the reliance of societies with regards to ecological systems and functions, as well as biodiversity. Disservices stand for the opposite of ecosystem services.

Introduction

Ecosystem services correspond to a modern way to figure out the role of the environment for human society, in particular in economics through the valuation of ecosystem goods and services (Gómez-Baggethun et al. 2010; Braat 2016). These services have been promoted by increasing and complementary contributions both by policymakers and academic researchers towards their agenda to promote sustainable development, especially regarding the environmental issues at stake.

Over the past decades, landmark international initiatives to assess their social value range from the Millenium Ecosystem Assessment (MEA) (2005) and The Economics of Ecosystems and Biodiversity (TEEB) (2010) to the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) (Pascual et al. 2017). In addition, international initiatives to set a policy agenda include among others initiatives by countries or group of countries (e.g., the US White House to include ES in Federal Policy (October 7, 2015) or the future EU Biodiversity strategy 2020), or initiatives by international bodies (e.g., the UN-REDD program to design and support Reducing Emissions from Deforestation and Forest Degradation + (REDD+) actions to cope with deforestation).

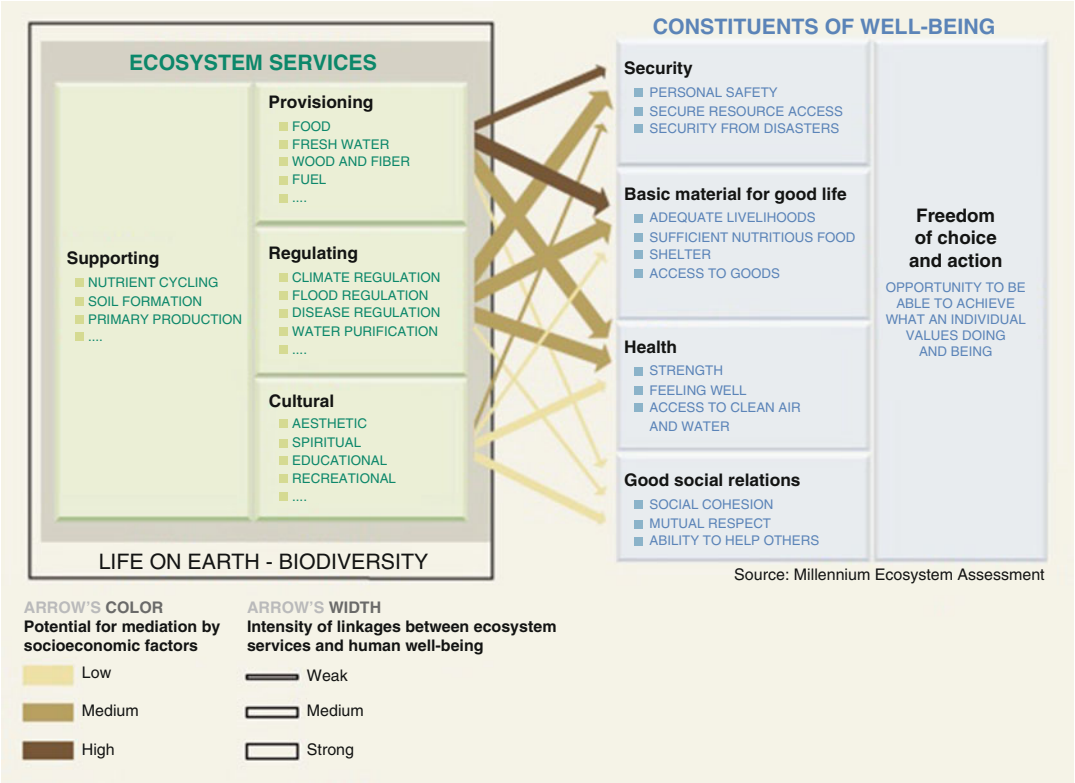
In terms of the history of economic thought, Mooney and Ehrlich (1997) date the modern history of the benefits got by human people from the ecosystems from the publication of Georges Perkins Marsh's book entitled *Man and Nature* in 1864. The concept of ecosystem service appeared effectively in the late 1970s (Ehrlich and Ehrlich 1981; Ehrlich and Mooney 1983), and the word service has been since used to reveal the benefits from ecological systems and functions in a utilitarianism framework, in order to highlight their social significance. According to Gómez-Baggethun et al. (2010), the role of Nature and then ecosystem services for human society depends on the considered economic perspective.

Firstly, Nature's benefits switch from solely use values in classical economics to exchange values in neoclassical economics. Secondly and more recently, in the environmental and resource economics literature, Nature's benefits are monetizable and exchangeable services, while in the ecological economics literature, there are controversies on monetization and commodification of Nature's benefits. More precisely, the latter distinction depends on the status granted to the natural capital and the extent of the literature at stake, from the inclusion of market failures to social ecological economics.

Ecosystem Service Classification and Valuation

Many ecosystem service classifications have emerged over the past three decades with regards to the contribution of different types of ecosystems to human well-being (MEA 2005; TEEB 2010). Following the revision of the System of Environmental-Economic Accounting (SEEA) under the umbrella of the United-Nations (UN), the Common International Classification of Ecosystem Services (CICES) was then developed to provide a hierarchically consistent and science-based classification (hosted by the European Environment Agency (EEA)).

Classifications of ecosystem services traditionally start from so-called supporting services that literally support other services mainly divided in provisioning services, regulating services and even maintenance services, and information services. Those services then contribute to well-being in providing security, basic material for good life, health, and social relations, and to a larger extent have public good characteristics. Figure 1 sums up these relationships in showing how ecosystem services and human well-being are linked from MEA (2005). We may underline that the status of

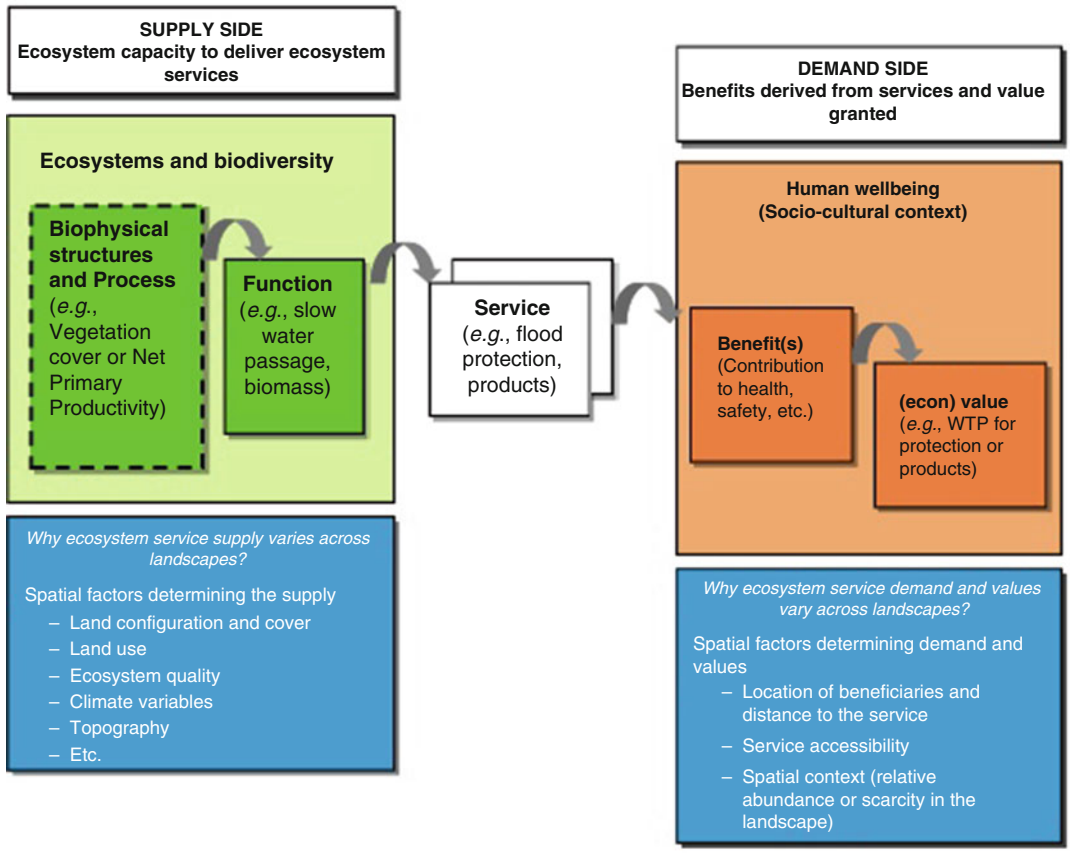


Ecosystem Services, Fig. 1 Relationships between ecosystem services and human well-being (from MEA (2005))

biodiversity remains debated among scholars as this can either be a support service through habitats for species or a single service as a whole. Moreover, relatively new contributions in the literature stress that multiple services are generated at the same time and consequently that this is suitable to use rather the concept of bundle of ecosystem services (Raudsepp-Hearne et al. 2010; de Groot et al. 2012).

Once classified and described with regards to an according typology, ecosystem services have to be valued to assess their contribution to society. Seminal papers in the literature range from methodological papers (de Groot et al. 2002) to effective value assessment even towards global values (Costanza et al. 1997, 2014; de Groot et al. 2012). Values are traditionally separated into ecological value, socio-cultural value, and economic value (de Groot et al. 2002). The economic value is then linked to use and non-use values and is measured in using economic valuation techniques either through direct market or indirect market/nonmarket valuation – avoided cost, replacement cost, travel cost, hedonic pricing, contingent valuation, and more recently choice experiment valuation – to derive willingness to pay (WTP) or willingness to accept (WTA) for the availability of these services.

Recent contributions highlight also the spatial dimension of ecosystem services and how this impacts their value (de Groot et al. 2012; Tardieu et al. 2015), either at the local scale or at the global scale through land use scenarios (Gascoigne et al. 2011). Consequently, the number, the location, and the characteristics of the beneficiaries may generate distance decays (Bateman et al. 2006; Tardieu et al. 2013). Figure 2 illustrates the spatial



Ecosystem Services, Fig. 2 Spatial dimension of ecosystem services: relationship between supply, demand and values (from Tardieu et al. (2013))

dimension of ecosystem services and therefore the overall relationship between supply, demand, and values (Tardieu et al. 2013) adapted from Haines-Young and Potschin (2010).

Environmental Law, Rules, and Incentives

Following the assessment of the significance of ecosystem services and their contribution to social welfare, the step forward consists in their inclusion in decision-making (Daily et al. 2009). Until the early 2000s, ecosystem services were indeed poorly included in cost-benefit analyses. The goal is then to integrate ecosystem services in landscape planning (de Groot et al. 2010) and more precisely strategic environmental planning and assessment (Kumar et al. 2013; Tardieu et al. 2015), to ensure the protection of the ecosystems. To this aim, environmental law can play a crucial role in allowing economic information to be used in decision-making.

According to Salzman (1998), environmental law can help for a better understanding of ecosystem services provision and then to assess their value by the creation of information markets. Indeed, the contribution of the economic analysis can take place only through a legal framework of rules and incentives. For example, the clean water acts regulation, the Code of Federal Regulations or the National Oceanic and Atmospheric Administration (NOAA) in the USA have created over the years information markets to gather data, to provide environmental impact statements, and to allow damages valuation on wetlands (Salzman et al. 2001). Moreover, the European Commission is currently preparing a Directive on the No Net Loss (NNL) of biodiversity and ecosystem services that can help to reach this aim (Wende et al. 2018).

Regarding the rules and incentives in environmental law, one could consider the following principles according to Salzman's "Five P's" (2005, 2013). These principles are respectively prescription, (financial) penalty, persuasion, property (rights), and payment. To sum up, prescription refers to the legal framework ranging from the command-and-control regulation to the expressive function of law with regards to social norms. Financial penalties alter behaviors through financial signals (e.g., a tax mechanism). Persuasion refers to information instruments towards stakeholders regarding suitable management practices and behaviors. Property rights embody the allocation of private rights or (transferable) use rights. Finally, payment often refers to positive monetary incentives (e.g., a subsidy mechanism). These principles are complementary and are applied differently according to the ecosystems and the services at stake.

Considering the property rights and the payment principles, this contributes to the creation of market-based instruments to protect ecosystem services (Salzman 2005; Miteva et al. 2012) and to a larger extent to the implementation of payments for ecosystem services (PES) (Alix-Garcia and Wolff 2014). PES reward the landholders either to enhance their activities in favoring the provision of ecosystem services or to cover their opportunity costs in giving up their activities if these are harmful for the ecosystem services in question. PES are increasingly used in the forestry sector regarding water purification and supply or carbon storage for example (Delacote et al. 2016). Note that there is currently an increasing literature to assess the successes and failures of PES systems.

Last, brand new mechanisms to protect ecosystem services are biodiversity offsets that take place under legal constraints. Biodiversity offsets are activities that provide measurable and additional ecological gains that are equivalent to the ecological losses in an impacted area. Biodiversity offsets operate even as a market of "mitigation credits" under the control of regulators if we take the example of wetland mitigation banking in the USA (Vaissière and Levrel 2015). In this case, exchanges between gains and losses are implemented to achieve no net loss of wetland and services protection, under the supervision of the United States Army Corps of Engineers (USACE) and the United States Environmental Protection Agency (USEPA).

Cross-References

- ▶ Non-Market Valuation
- ▶ Public Goods

References

Alix-Garcia J, Wolff H (2014) Payment for ecosystem services from forests. Ann Rev Resour Econ 6(1):361–380

Bateman IJ, Day BH, Georgiou S, Lake I (2006) The aggregation of environmental benefit values: welfare measures, distance decay and total WTP. Ecol Econ 60:450–460

Braat LC (2016) Ecosystem services. Oxford Research Encyclopedia of Environmental Science. https://doi.org/10.1093/acrefore/9780199389414.013.4

Costanza R, d'Arge R, de Groot R, Farber S, Grasso M, Hannon B, Limburg K, Naeem S, O'Neill RV, Paruelo J, Raskin RG, Sutton P, van den Belt M (1997) The value of the World's ecosystem services and natural capital. Nature 387:253–260

Costanza R, de Groot R, Sutton P, van der Ploeg S, Anderson SJ, Kubiszewski I, Farber S, Turner RK (2014) Change in the global value of ecosystem services. Glob Environ Chang 26:152–158

Daily GC, Polasky S, Goldstein J, Kareiva PM, Mooney HA, Pejchar L, Ricketts TH, Salzman J, Shallenberger S (2009) Ecosystem services in decision making: time to deliver. Front Ecol Environ 7(1):21–28

de Groot RS, Wilson MA, Boumans RJM (2002) A typology for the classification, description and valuation of ecosystem functions, goods and services. Ecol Econ 41:393–408

de Groot RS, Alkemade R, Braat L, Hein L, Willemen L (2010) Challenges in integrating the concept of ecosystem services and values in landscape planning, management and decision making. Ecol Complex 7:260–272

de Groot RS, Brander L, van der Ploeg S, Costanza R, Bernard F, Braat L, Christie M, Crossman N, Ghermandi A, Hein L, Hussain S, Kumar P, McVittie A, Portela R, Rodriguez LC, ten Brink P, van Beukering P (2012) Global estimates of the value of ecosystems and their services in monetary units. Ecosyst Serv 1:50–61

Delacote P, Robinson EJZ, Roussel S (2016) Deforestation, leakage and avoided deforestation policies: a spatial analysis. Resour Energy Econ 45:192–210

Ehrlich PR, Ehrlich AH (1981) Extinction: the causes and consequences of the disappearance of species. Random House, New York

Ehrlich PR, Mooney HA (1983) Extinction, substitution, and ecosystem services. Bioscience 33(4): 248–254

Gascoigne WR, Hoag D, Koontz L, Tangen BA, Shaffer TL, Gleason RA (2011) Valuing ecosystem and economic services across land-use scenarios in the Prairie Pothole region of the Dakotas, USA. Ecol Econ 70:1715–1725

Gómez-Baggethun E, de Groot R, Lomas PL, Montes C (2010) The history of ecosystem services in economic theory and practice: from early notions to markets and payment schemes. Ecol Econ 69(6):1209–1218

Haines-Young R, Potschin M (2010) The links between biodiversity, ecosystem services and human well-being. In: Raffaelli D, Frid C (eds) Ecosystem ecology: a new synthesis. BES ecological reviews series. Cambridge University Press, Cambridge

Kumar P, Esen SE, Yashiro M (2013) Linking ecosystem services to strategic environmental assessment in development policies. Environ Impact Assess Rev 40:75–81

Millenium Ecosystem Assessment (MEA) (2005) Ecosystems and human well-being: synthesis. Island Press, Washington, DC

Miteva DA, Pattanayak SK, Ferraro PJ (2012) Evaluation of biodiversity policy instruments: what works and what doesn't? Oxf Rev Econ Policy 28(1):69–92

Mooney H, Ehrlich P (1997) Ecosystem services: a fragmentary history. In: Daily GC (ed) Nature's services. Island Press, Washington, DC, pp 11–19

Pascual U et al (2017) Valuing nature's contributions to people: the IPBES approach. Curr Opin Environ Sustain 26:7–16

Raudsepp-Hearne C, Peterson GD, Bennett EM (2010) Ecosystem service bundles for analyzing tradeoffs in diverse landscapes. Proc Natl Acad Sci USA (PNAS) 107(11):5242–5247

Salzman J (1998) Ecosystem services and the law (Editorial). Conserv Biol 12(3):497–498

Salzman J (2005) Creating markets for ecosystem services: notes from the field. N Y Univ Law Rev 80:870–961

Salzman J (2013) Teaching policy instrument choice in environmental law: the Five P's. Duke Environ Law Policy Forum 23(2):363–373

Salzman J, Thompson BH, Daily GC (2001) Protecting ecosystem services: science, economics and law. Stanf Environ Law J 20:309–332

Tardieu L, Roussel S, Salles J-M (2013) Assessing and mapping global climate regulation service loss induced by terrestrial transport infrastructure construction. Ecosyst Serv 4:73–81

Tardieu L, Roussel S, Thompson JD, Labarraque D, Salles J-M (2015) Combining direct and indirect impacts to assess ecosystem service loss due to infrastructure construction. J Environ Manag 152:145–157

The Economics of Ecosystems and Biodiversity (TEEB) (2010) In: Kumar P (ed) The economics of ecosystems and biodiversity: ecological and economic foundations. Earthscan, London/Washington, DC

Vaissière A-C, Levrel H (2015) Biodiversity offset markets: what is this really about? An empirical approach of wetlands mitigation banking. Ecol Econ 110:81–88

Wende W, Tucker G, Quétier F, Rayment M, Darbi M (2018) Biodiversity offsets - European perspectives on no net loss of biodiversity and ecosystem services. Springer. https://doi.org/10.1007/978-3-319-72581-9

Efficiency

Fatih Deyneli
Department of Public Finance, Pamukkale University, Denizli, Turkey

"An effective judiciary is predictable, resolves cases in a reasonable time frame, and is accessible to the public."
Marina Dakolias 1999 Court Performance around the World: A Comparative Perspective

Definitions of Efficiency in Economics and in Law

The economic analysis of law relies upon microeconomic principles. Efficiency, a significant concept within microeconomic theory, is one of the key concepts in the economic analysis of law. There are different types of efficiency in economic literature. These types are defined as follows.

Productive Efficiency: The very first answer to the question "what is efficiency?" may be the ratio of the work done or output obtained by a certain amount of input. This is a state of maximum output using a certain source and technology during production. In other words, we can accomplish maximum output with the available source and technology we have. If the production with the sources and technology at hand is lower than maximum output, the state of inefficiency comes into picture (Arnold 2011).

Pareto Efficiency: Created by nineteenth century economist and sociologist Vilfredo Pareto (1848–1923), this term is a beneficial criterion in comparing the output from different economic institutions. If resources can be allocated evenly in which it is impossible to make any one individual better off without making at least one individual worse off, a Pareto improvement can be obtained. If allocation leads the way to Pareto improvement, this situation is called Pareto inefficient. In this situation, economy cannot have all that it could obtain from its resources. If we cannot find a way to make any one individual better off without making at least one individual worse off, this situation is called Pareto efficient. In this situation, economy can have all it could obtain from its resources (Varian 2010; Besanko and Braeutigam 2011).

Kaldor-Hicks Efficiency: The application of Pareto efficiency concept to justice leads to a real problem. Firstly, there is one winner and one loser in justice, which originates from the compulsory characteristic of justice (Dimock 2005). Secondly, Pareto efficiency is also applicable to free market rules. Justice services, necessary for the implementation of law, are public commodities. Therefore, Pareto efficiency needs transformation before being used in the economic analysis of law.

Kaldor and Hicks transformed Pareto efficiency so that it can be added to the area of law which has a compulsory characteristic. Kaldor-Hicks efficiency criterion can be defined as: "a change is an improvement by Kaldor-Hicks criterion if the gainer value their gains more highly than the losers their losses" (Mathis 2009). This is basically a cost-benefit analysis. In cost-benefit analysis, when benefit exceeds cost, the project is commenced so it implies that winners may regulate losers (Cooter and Ulen 2011).

Justice System Efficiency

According to Posner, justice system has two expenditures, which are fault expenditure and direct expenditure. Fault expenditure is the failure of justice system in the distribution and other social functions. Direct expenditure is the cost of time, stationery, phone, and office work spent by judges, lawyers, and others. Posner states that fault expenditures are as real as direct ones and the economic aim should be to lower these two expenditures. The aim of justice system should be to lower both fault and direct expenditures (Posner 1973).

European Commission for the Efficiency of Justice defines justice system as a system composed of input, production, and output. These inputs, productions, and outputs within this system are summarized in Table 1 (Albers 2003). The demand side of justice system involves the

Efficiency, Table 1 Input, production, and output of justice system according to CEPEJ

Input	Production	Output
Number of courts	Average time of criminal cases	Number of resolved criminal cases
Court budget		
Information budget of courts	Average time of commercial cases	Number of resolved commercial cases
Numbers of professional judges		
Numbers of members of the jury	Average time of divorce cases	Number of resolved divorce cases
Numbers of deputy judges		
Numbers of justice staff	Average time of social security cases	Number of resolved social security cases
Numbers of other staff		
Average judge salary	Average time of immigrant cases	Number of resolved immigrant cases
Average justice staff salary		
Number of criminal cases	Average time of fiscal cases	Number of resolved fiscal cases
Number of commercial cases		
Number of divorce cases	Average time of labor cases	Number of resolved labor cases
Number of social security cases		
Number of fiscal cases		
Number of labor cases		

Source: Albers 2003

processing of incentives and constraints in lawyers' and parties' behavior (Rosales-Lopez 2008). The supply side of justice system involves factors such as the budget; number of judges, prosecutors, and assistant staff and their working time; and technology (Buscaglia and Ulen 1997).

The demand side of justice system and fault expenditure is mostly related to Kaldor-Hicks efficiency concept, while the supply side of justice system and direct expenditures is mostly related to the concept of productive efficiency.

Common Efficiency Indicators

Economic analyses have also begun to be used in the economic analysis of law. Quantitative measurements such as numbers of incoming cases, disrupted cases, and resolved cases and numbers of crime and accident have also gained importance in these analyses (Posner 2006). Due to different judicial systems in different countries, there are some differences among indicators. Despite these differences, common indicators about the efficiency, quality, and performance of justice services in international level have been used as reference point in many studies (Dakolias 1999). Developed for the measurement of justice system efficiency, these indicators benefit in objectifying the concept of justice. By this way, the concept of justice becomes appropriate for economic and econometric analyses.

The most commonly used one of these indicators is clearance rate. International institutions make use of other indicators to analyze the efficiency of justice services. Table 2 provides the websites of

Efficiency, Table 2 International institutions working on the efficiency of justice system

Name of institution	Web page
OECD	http://www.oecd.org
World Bank	http://www.worldbank.org/
European Commission	http://ec.europa.eu/justice/
The European Commission for the Efficiency of Justice (CEPEJ)	http://www.coe.int/t/dghl/cooperation/cepej/default_en.asp
National Center for State Courts	http://www.ncscinternational.org/
International Consortium for Court Excellence	http://www.courtexcellence.com/

international institutions for more indicators and analysis of these indicators on country basis.

Clearance rate is calculated by dividing the number of crimes that are "cleared" (a charge being laid) by the total number of crimes recorded. It may also be defined as the response of justice system to the clearance rate and demand in justice system – applications to the justice system. If clearance rate is more than 100%, courts begin to deal with the cases remained from the previous year. If clearance rate is less than 100%, there will be cases unresolved (Dakolias 1999). Clearance rate is measured as follows (CEPEJ 2012):

$$\text{Clearance rate} = \frac{\text{Resolved cases in a period}}{\text{Incoming cases in a period}} \times 100$$

References

Albers P (2003) Evaluating judicial systems "a balance between variety and generalisation". CEPEJ 12, Strasbourg

Arnold R (2011) Microeconomics, 11th edn., South-Western Cengage Learning, USA

Besanko D, Braeutigam R (2011) Microeconomics, 4th edn., NJ: Wiley, USA

Buscaglia E, Ulen T (1997) A quantitative assessment of the efficiency of the judicial sector in Latin America. Int Rev Law Econ 17:275–291

CEPEJ (2012) CEPEJ report evaluating European judicial systems – 2012 edition (2010). CEPEJ studies no 18, Belgium

Cooter RB, Ulen T (2011) Law and economics, 6th edn., Pearson, USA.

Dakolias M (1999) Court performance around the world: a comparative perspective. World Bank, 83 WTP no 430, Washington, D.C

Dimock S (2005) Law and economics. Classic readings and Canadian cases in the philosophy of law. Prentice Hall, Toronto

Mathis K, Shannon D (trans) (2009) Efficiency instead of justice? Searching for the philosophical 87 foundations of the economic analysis of law. Springer Science +Business Media B.V, New York

Posner RA (1973) An economic approach to legal procedure and judicial administration. J Legal Stud 2(2):399–458

Posner RA (2006) A review of Steven Shavell's foundations of economics analysis of law. J Econ Lit 44(2):405–414

Rosales-Lopez V (2008) Economics of court performance: an empirical analysis. Eur J Law Econ 25:231–251

Varian HR (2010) Intermediate microeconomics, 8th edn. W.W. Norton, New York

Efficiency, Types of

Henk J. ter Bogt
University of Groningen, Groningen, The Netherlands

Abstract

Efficiency is an important concept in the field of law and economics. The precise meaning of the term efficiency is not unequivocal and depends on the context in which it is used. In practice, efficiency often relates to economic efficiency. However, such forms as social and political efficiency can also be very important for the continuity of a society or an organization.

Efficiency types in organizations and society: Different kinds of measures which all indicate the optimal use of specific material or immaterial scarce resources in organizations and society.

Introduction

Efficiency is a well-known word in everyday language. It is also a core concept in the law and economics literature, a field which deals with the economic effects of legal rules and constraints and the way in which these rules and constraints can contribute to "efficiency." To put it in more general terms, it could be claimed that the traditional emphasis of the law and economics literature has been on finding efficient legal rules and structures, i.e., instruments which promote economic efficiency. This objective implies finding an optimal allocation of the resources available and realizing a maximum quantity of outputs via these means. Laws and rules can play a role here (Coase 1960; Calabresi 1961). In principle people are supposed to act rational in the law and economics literature, which implies that based on their economic rationality, they strive for economic efficiency. In later additions to the law and economics literature, however, efficiency is used in broader terms,

referring to more aspects than only economic efficiency.

The specific focus on "efficiency" might suggest that the precise meaning of the concept is always clear. However, in economics different types of efficiency can be distinguished. And in later elaborations of the law and the economics literature, even more types have been added, such as social efficiency and political efficiency, each with their own specific definitions. All of these definitions may refer to "maximizing" something, i.e., an output, by using a minimum amount of "inputs." What has to be maximized, and what the inputs and outputs are, varies and relates to the specific angle and subject area chosen. This optimization could relate, for example, to economic welfare, but also to public benefit (which can include immaterial and nonmonetary elements). Furthermore, in the later additions to the law and economics literature, human behavior has received increasing attention (see, e.g., Jolls et al. 1998; Korobkin and Ulen 2000). This has been done, for example, via concepts such as bounded rationality (people may want to choose the best alternative available to them, but they can only acquire full knowledge of a limited number of alternatives) and bounded self-interest (although people may be focussed on their own interests and utility, this does not necessarily imply that they maximize utility without caring about others). Despite these refinements, however, in practice the ultimate focus of much of the law and economics literature still seems to be on the realization of optimal economic efficiency.

Below, first, the basic concepts of inputs and outputs will be introduced, followed by a discussion of the more traditional economic efficiency concept and the elements in which it can be subdivided. Next, some other types of efficiency will be discussed briefly and a conclusion will be presented.

Inputs and Outputs

Some of the concepts mentioned above require some further explanation, because they are the basic components in the analysis of efficiency. The various forms of efficiency all deal with inputs and outputs. In a production process, examples of the *inputs* required to produce a certain output, i.e., a good or a service, are direct and indirect labor hours, capital, materials, machine hours, ICT, housing, knowledge, and transport. These inputs can also be called resources. Together, the inputs are the costs made in the production of the outputs. The *outputs* are the "products" produced, for example, the goods and services, which are the result of a production process in an organization. So the production process transforms the inputs into outputs (which may lead to certain effects, i.e., outcomes considered desirable by an individual or society, resulting, e.g., in utility maximization). In order to be able to measure and compare the different inputs, they generally have to be expressed in monetary values.

Economic Efficiency and Its Dimensions

The term efficiency as commonly used generally refers to the ratio between the inputs employed and the outputs realized. More precisely, it refers to the maximization of output produced by a unit of input. If more output can be produced per unit of input, the efficiency increases. Linking outputs to inputs in this way could also be called measuring the productivity. Depending on how much output is produced by a unit of input, the common term "efficiency" can also indicate organizations as being less or more efficient. Therefore, the general criterion based on which different organizations are compared is their "efficiency." However, in the academic economic and law and economics literature, many authors define "efficient" in more specific and absolute terms.

When the term efficiency is used in the field of law and economics, it generally refers to the so-called economic efficiency, which can be subdivided into two types: productive or technical efficiency and allocative efficiency. These categories together form the overall economic efficiency (Coelli et al. 1998, pp. 3–5).

In an economic context, the first type – productive or technical efficiency – is defined as

follows: "efficiency is the difference between the actual input/output ratio and the ideal ratio" (Jackson 2011, p. 16), all other things – such as the production technology and the quality of the output – remaining equal. So the ideal ratio is considered to be the result of the optimal possible use of a certain quantity of input, and in that situation the input yields the maximum quantity of output attainable on the basis of this input (i.e., the "production frontier," as it is labeled by economists, has been reached). In economics, this type of efficiency is called *productive* or *technical efficiency* (terms which can be considered as synonyms). Productive (or technical) efficiency relates to an organization's production costs. It is achieved when the organization's outputs are produced at minimum average costs (i.e., only if this point of minimum average costs is achieved can one speak of an organization being "efficient").

Although organizations and/or society can be productively efficient, this does not automatically imply that the products produced are valued by the citizens/consumers. The other dimension of economic efficiency, called the *allocative efficiency*, therefore addresses the question whether the "right products" are produced. If a company would produce steam engines in a productively efficient way whereas nobody needed these items, a major problem could arise concerning using the resources for this production process. Allocative efficiency indicates whether the resources in society are allocated in an optimal way, given a particular behavioral assumption, for example, utility maximization (but the behavioral assumption could also be, e.g., profit maximization or cost minimization). This means that allocative efficiency concentrates on whether a society's resources (inputs) are allocated to those domains of activity and output which bring the citizens/consumers the maximum utility (where in this context utility is supposed to be related to the availability of economic goods). Expressed in more technical economic terminology, allocative efficiency refers to a situation where the price which consumers are willing to pay for a product (this price indicates the utility or value attached to the product) is equal to the cost of the resources used for realizing this product (i.e., the marginal cost of the product). In an alternative definition allocative efficiency indicates the situation in society in which the available resources (inputs) are used in such a way that it would not be possible to increase one person's utility without decreasing that of another. In economics, the latter example is also called a Pareto-efficient or Pareto-optimal allocation.

A concept related to the productive or technical efficiency category is the so-called *X-efficiency*. Being X-efficient means that an organization employs its internal resources in an optimal manner, i.e., in such a way that the average costs are reduced to a minimum. However, if the competition in the market is imperfect, for example, because of a monopoly, a situation of X-*in*efficiency may exist, also on a longer-term basis (Leibenstein 1966). Here the organization is not capable of achieving productive or technical efficiency and produces its outputs at higher costs per unit than technically considered necessary.

The above depicts the most frequently used and most important dimensions of economic efficiency. However, apart from the strictly economic forms of rationality, there are also other variants relevant in practice, which are not necessarily of a higher or lower order. Next, an outline is given of two of them.

Social Efficiency

Although social efficiency relates to an assessment of the welfare in a society in a mainly economic sense, it has more recently also been used in connection with all kinds of noneconomic topics (e.g., social, political, and cultural) which can play a role in and influence the public benefits. When used in this broader context, social efficiency refers to the maximization of certain "outputs" in society (e.g., the contribution to an objective such as sustainability or social justice) via the inputs available (see, e.g., Lefeber and Vietorisz 2007).

The more traditional use of the term social efficiency, however, goes back to Coase (1960). In that context *social efficiency* indicates the maximization of the social welfare, which means

a maximization of the revenues obtained. This maximization of revenues suggests an economic focus. The revenues are considered to be maximized if the inputs and outputs are used in such a way that the social cost of production is equal to the social benefit (or, in more technical economic terminology, if the social marginal benefit of the consumption of the output is equal to the social marginal cost). If expressed in this way, social efficiency seems to be closely linked to allocative efficiency. However, the concept of social efficiency includes paying attention to the existence of the so-called externalities (and markets which are not perfectly competitive). When focussing on welfare at the level of society, the effect of production on different individuals or groups is considered with the aim to maximize the social welfare, i.e., attaining social efficiency.

As social efficiency and "externalities" are important in much of the law and the economics literature, they are dealt with in some more detail. External effects relate to the influence of the activity or action of an individual or organization on other individuals or organizations. Such an influence could be, for example, the pollution caused by the activities of a factory in the gardens of the people living in the vicinity. This effect, an externality, should be taken into consideration when social welfare has to be determined. In the case of the factory, elements should be taken into account like the negative effect of the production activities on the living environment of the people in the surrounding area and the value of their property. In principle, the factory could negotiate about a compensation for such negative effects with each individual neighbor. However, if the factory would have to negotiate with all its neighbors individually, its "transaction costs" may rise to an unacceptable level (in this case the transaction costs could include, e.g., the costs involved in searching for alternatives to avoid or compensate the pollution, negotiations, and the monitoring of the agreements made). Here certain legal rules which prescribe how to act in the case of an externality (and what kind of negative effects has to be accepted by one or both of the parties) could be helpful in reducing the cost of making contractual arrangements (and in reducing transaction costs) and approaching an optimal economic result for society (Coase 1960; Calabresi 1961).

Although social efficiency as developed on the basis of the ideas of Coase and Calabresi mostly refers to economic welfare, Coase (1960, p. 43) suggested that the analysis "of different social arrangements for the solution of economic problems should be carried out in broader terms than" only the economic context (i.e., it should be broader than only the "comparison of production values as measured by the market"). If such a broader interpretation of the concept of efficiency is adopted, it seems to be very similar to the abovementioned maximization of social benefits.

Political Efficiency

In a political environment, the goal function, i.e., ter what has to be maximized (the "outputs"), may differ from that in a "strictly economic" context, which also applies to the inputs concerned. This could particularly be the case when special interest groups try to influence political decision-making (Buchanan and Tullock 1974, pp. 284–287, 302–305). Since the law and economics literature considers both the effects of legal rules and the way in which they develop and are ultimately formulated – which is basically a political process – several authors have also included the political context in their work.

A politician who has to decide about the content of a law could be interested in its effects in terms of both economic and social efficiency. Taking economic and social efficiency into account can be important in avoiding dissatisfaction from within society. However, the politician may also be interested in the effect of the legal rules on his/her voters and the number of votes he/she and his/her party will receive in the next elections, i.e., political efficiency (Buchanan 1993; Gavious and Mizrahi 2002). From the perspective of political rationality, i.e., to survive as a political party or a politician, a rational approach would be to try to maximize the electoral support by using minimal "efforts." *Political efficiency*

could then be expressed "as a ratio of the amount of 'effort' (including money and other production factors) made by politicians to the amount of electoral support gained by means of the politicians' policies" (ter Bogt 2003, p. 180). The elements on which the politician's efforts should focus can vary from case to case. This is because elements such as the public interest and social welfare are generally not that clearly defined in practice and can include many different components, which can also differ among citizens (who are the voters in elections). Furthermore, the political support of voters may not only depend on the economic efficiency of the government policy but also on other factors, for example, equal treatment, a focus on sustainability, and social justice (see also Wildavsky 1966, pp. 308–309).

To Conclude

In principle, depending on the specific context, the type of rationality may vary, and herewith the relevant type and definition of "efficiency." Economic and law and economics literature were traditionally mainly focussed on economic rationality and efficiency. However, social efficiency and the related concept of social rationality have now also become important elements in the law and economics literature. Social efficiency as it is mostly used in practice is closely linked to economic welfare in society. But it can also include "noneconomic" elements, such as sustainability or social justice, and pertain to the broader level of public benefits. The idea of political rationality and political efficiency applies more specifically to political environments (although it seems that in practice "politics" can play a role in any organizational context). The political efficiency idea is partly related to Granovetter's suggestion that in order to better understand economic activity and the organizational arrangements chosen, it is necessary to take both economic and social relations into consideration (Granovetter 1985, pp. 490–491).

In this way the law and economics literature and the concept of efficiency have gradually been extended. The literature no longer only focusses on economic aspects but also includes behavioral elements, such as bounded rationality, as well as goals and outputs other than strictly economic ones. It now tries to include more elements from the "real world." This approach leads to different types of efficiency, playing a role in different contexts. As might have become obvious from the examples given above, the operationalization and proper measurement of the various concepts and types of efficiency in empirical research can be problematic, even in the case of economic efficiency. However, this difficulty does not imply that it would be a futile exercise to consider – and if otherwise not possible only in qualitative terms – the effects of changes in rules and structures on a specific type of efficiency in an organization or in society.

References

Buchanan JM (1993) How can constitutions be designed so that politicians who seek to serve "public interest" can survive and prosper? Constit Polit Econ 4(1):1–6

Buchanan JM, Tullock G (1974) The calculus of consent: logical foundations of constitutional democracy. University of Michigan Press, Ann Arbor

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

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

Coelli T, Prasada Rao DS, Battese GE (1998) An introduction to efficiency and productivity analysis. Kluwer, Boston/Dordrecht/London

Gavious A, Mizrahi S (2002) Maximizing political efficiency via electoral cycles: an optimal control model. Eur J Oper Res 141(1):186–199

Granovetter M (1985) Economic action and social structure: the problem of embeddedness. Am J Sociol 91(3):481–510

Jackson PM (2011) Governance by numbers: what have we learned over the past 30 years? Public Money Manag 31(1):13–26

Jolls C, Sunstein CR, Thaler R (1998) A behavioral approach to law and economics. Stanford Law Rev 50:1471–1550

Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. Calif Law Rev 88(4):1051–1144

Lefeber L, Vietorisz T (2007) The meaning of social efficiency. Rev Polit Econ 19(2):139–164

Leibenstein H (1966) Allocative efficiency vs. "X-efficiency". Am Econ Rev 56(3):392–415

ter Bogt HJ (2003) A transaction cost approach to the autonomization of government organizations. Eur J Law Econ 16(2):149–186
Wildavsky A (1966) The political economy of efficiency: cost-benefit analysis, systems analysis, and program budgeting. Public Adm Rev 26(4):292–310

Efficient Market

Michael Mitsopoulos[1] and Theodore Pelagidis[2]
[1]Brookings Institution, Athens, Greece
[2]Brookings Institution, University of Piraeus, Piraeus, Attica, Greece

Introduction

In this article, we deal with some pieces of evidence that are needed to explain the paradox of rapid GDP growth in the face of the dismal competitiveness and inefficiency of the Greek economy during 1995–2008. We show how Greece's economy structural inefficiencies – inefficiency of justice included – have hit the domestic economy, and we present their impact on the current turmoil of the economy. We offer a specific explanation of the current unfortunate state of the economy, and we briefly summarize the reforms already undertaken towards efficiency. We also suggest avenues of necessary reforms to increase efficiency and overcome the current crisis.

The Engines of Growth, 1995–2008

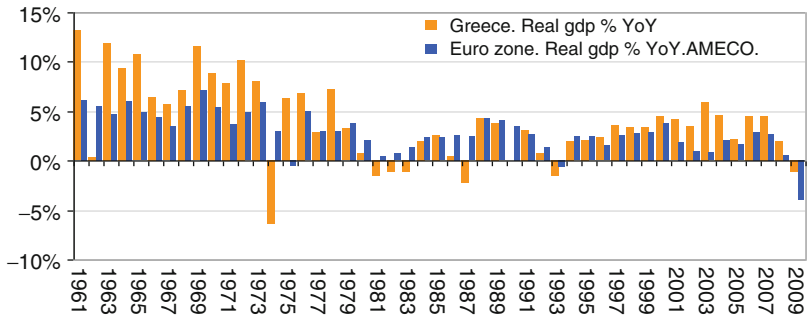
In Greece, certain positive developments led to the strong growth performance observed since the mid-1990s and up to 2008. Figure 1 shows how Greece clearly outperformed, since 1995–1996, the benchmark eurozone economy. However, it is absolutely crucial to look at the factors of "growth" to see why, at least in the great part, this was superficial, fragile, and not based on the improvement, the deepening, or the expansion of domestic production.

These developments include, primarily, the proper liberalization of the credit markets at the beginning of the 1990s, completed by the end of the 1990s. This was coupled with entry to the European Monetary Union. Combined these two developments lead simultaneously to macroeconomic stabilization and a steady increase of private credit after 2000. It has also to be stressed that the expansion of private credit replaced after the beginning of the 1990s the government deficit spending as the main way to finance the expansion of consumption in Greece, although data should be treated with caution. The most possible is that fiscal expansion reinforced private credit and private consumption expansion. As Fig. 2 shows by measuring demand injections to the GDP, the impact of these injections was important as a percentage of GDP for every year during a prolonged period that spans all the duration of Greece's strong performance.

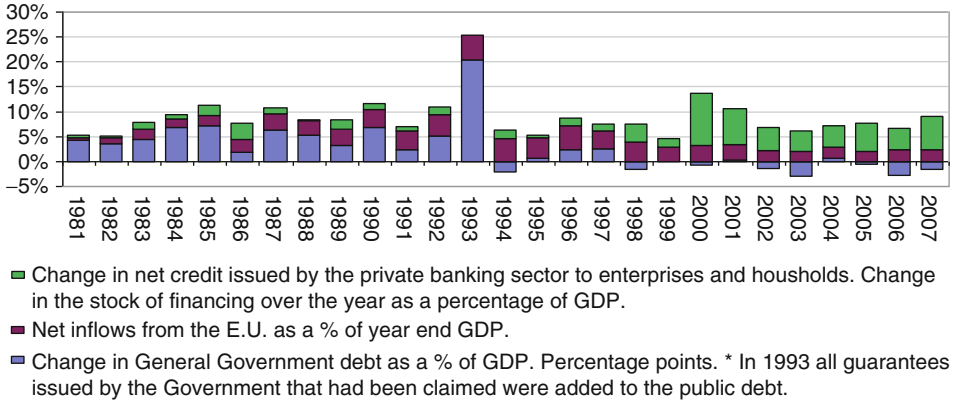
The contribution of the stabilization of the macroeconomic outlook of Greece in the wake of EMU accession towards the expansion of private credit was significant, as is shown by the rapid fall of interbank rates after 1998 (Fig. 3), which reflect also the decline in the rates offered by commercial banks to households and businesses. (It also brought a significant fall of the inflation differential of Greece with respect to the eurozone average during the same period). It can be seen clearly how the expansion of credit to households fuelled the growth of private consumption during the past years (Fig. 4), and only just the period preceding the completion of the infrastructure projects that were prepared to be ready for the 2004 Olympic Games, private consumption kept accelerating in spite of a lull in the explosive growth of private sector credit.

But this exception is easily explained by the peak in the investment growth rate during that time (Fig. 5).

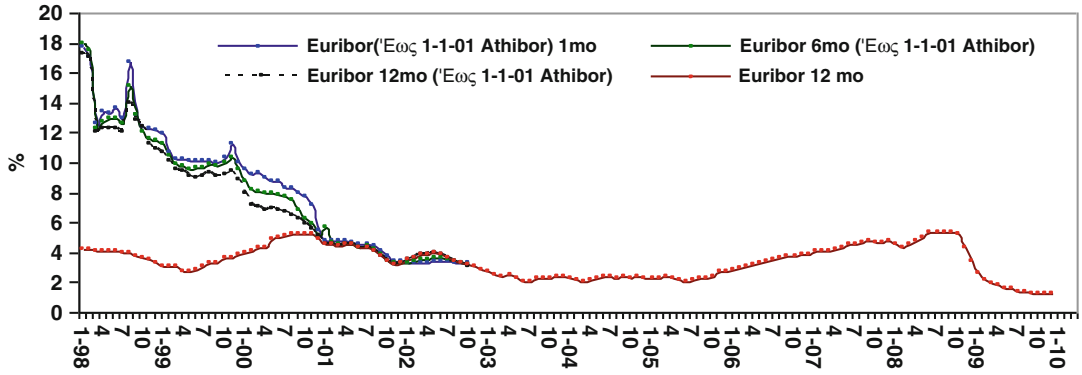
Besides the credit expansion, two other factors contributed significantly to Greece's growth performance during the 2000s. First is the shipping and tourism industry. These secure significant annual revenue inflows of about 25% of GDP that are added to the domestic demand and help to mitigate the huge trade balance deficit. Secondly, the fiscal stimulus given by the 2004 Olympic Games nourished through public borrowing



Efficient Market, Fig. 1 Real GDP (Source: AMECO)



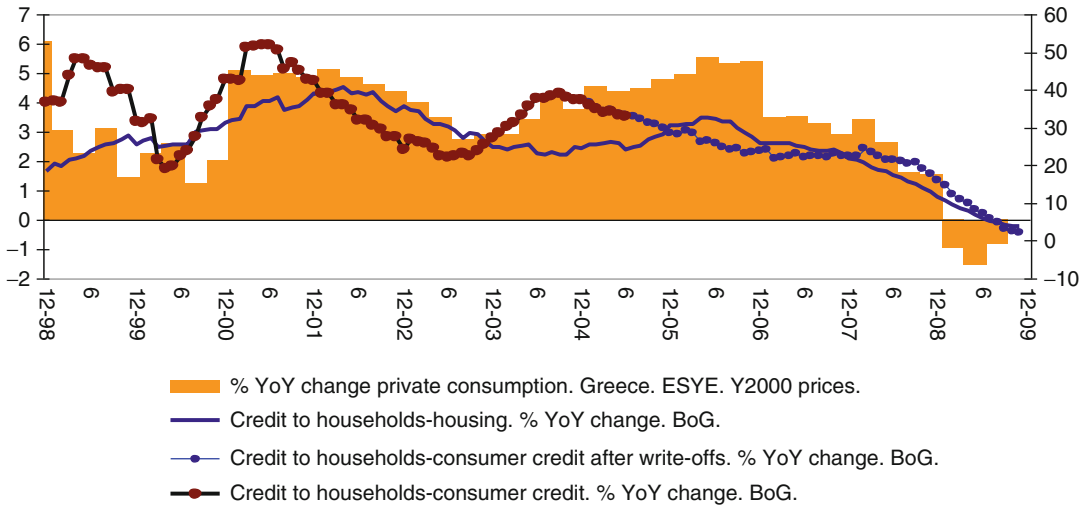
Efficient Market, Fig. 2 Demand injections (Source: Bank of Greece, Ministry of Finance, European Commission Budget, and EUROSTAT)



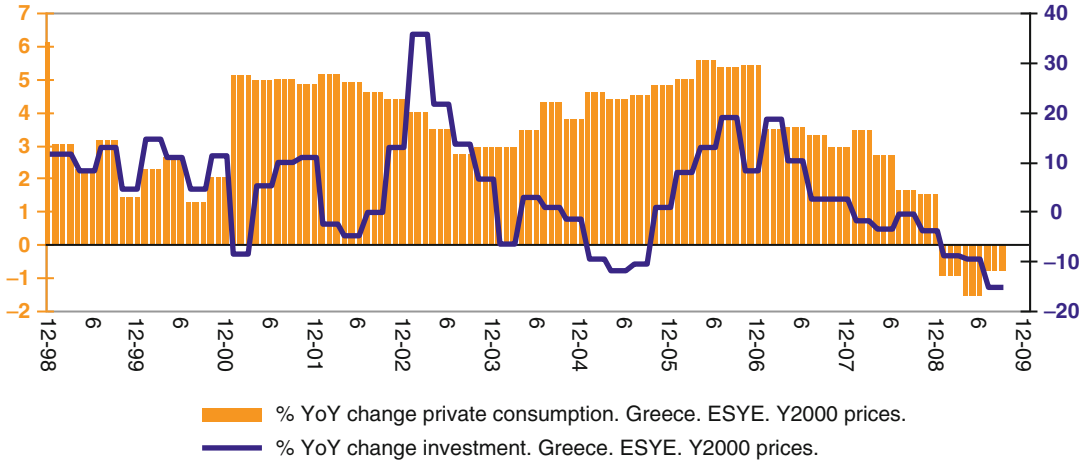
Efficient Market, Fig. 3 Interbank rates

and that led to the improvement of certain key infrastructure facilities.

The rapid increase of new investment, both public and private (Fig. 6), also demonstrates the impact of the infrastructure investment that was largely financed by the EU structural funds. Still, the rush into EU-financed infrastructure investment did not only contribute to investments and consequently to the creation of new jobs, as in the end many of these projects, when finished, actively boosted to some extent the productivity in the area surrounding Athens. The inflow of



Efficient Market, Fig. 4 % YoY change private consumption and Credit to households

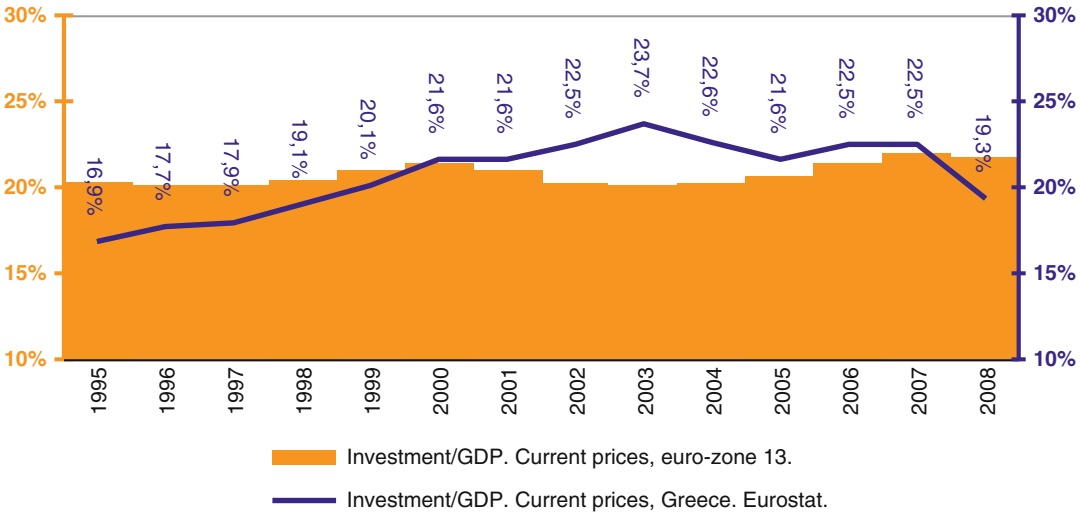


Efficient Market, Fig. 5 % YoY change private consumption & investment. Greece. ESYE. Y2000 prices

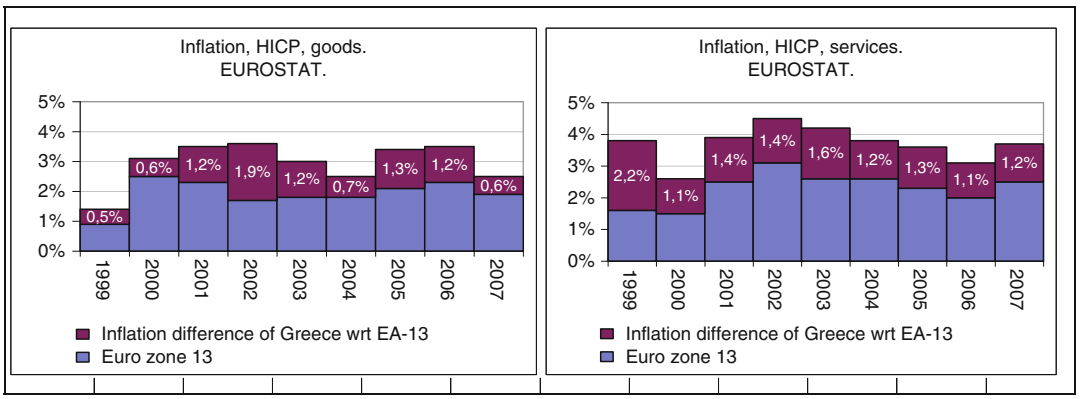
funds from the European Union, within the context of the European Union structural funds and the Common Agricultural Policy, also contributed largely to the improvement of key productivity enhancing infrastructure facilities.

The Facets of Low Competitiveness and Inefficiency

However, at the same time, a wide range of factors persisted in contributing towards the poor performance in certain other aspects of the Greek economy. The poor performance regarding competitiveness, to name just the most important one, is not only documented by numerous databases and surveys by international organizations and researchers but also by the persistent deficit of the current account in double-digit numbers (as a% of GDP), the persisting positive differential with the eurozone average inflation, and the unattractiveness of Greece to foreign direct investments that are practically zero (inflows minus outflows). Relatively recent research by



Efficient Market, Fig. 6 Investment as % of GDP



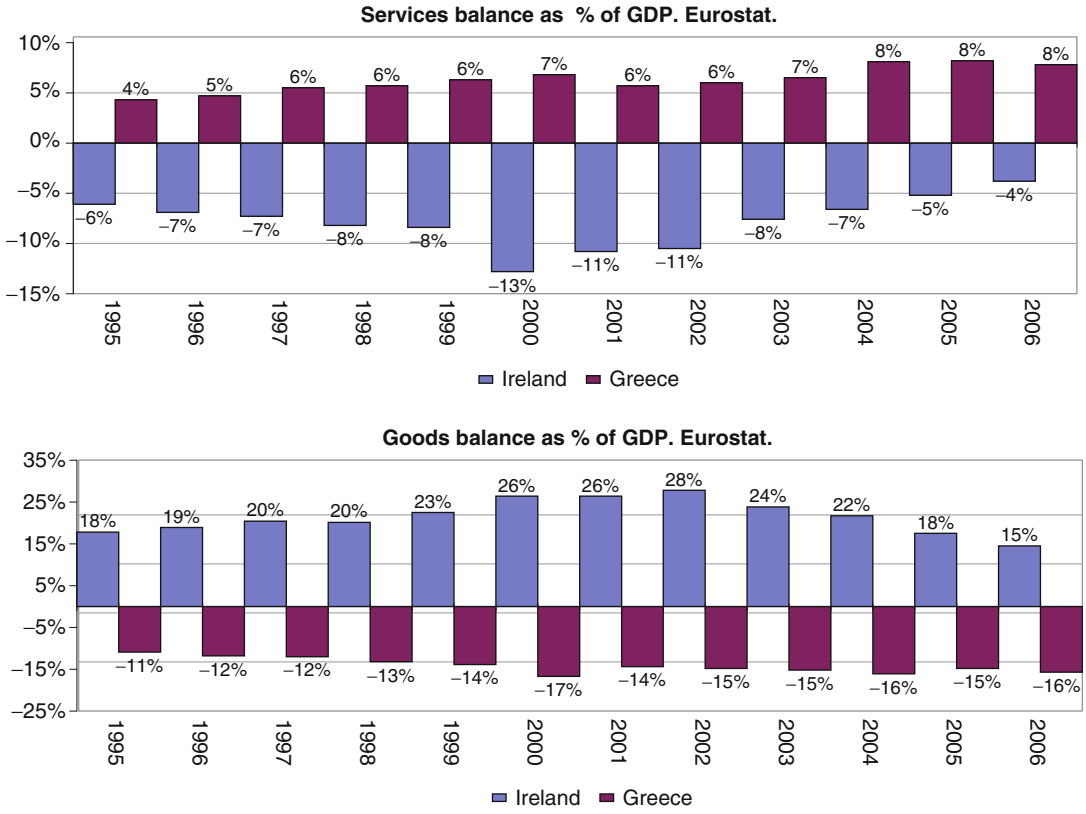
Efficient Market, Fig. 7 Inflation, HICP, goods & services. EUROSTAT

institutions like the OECD and the World Bank and a detailed presentation of numerous pieces of evidence indicate that the wide range of institutional weaknesses that prevail in Greece account, as a whole, for this dismal competitiveness performance.

The interesting part about the inflation differential of Greece with the eurozone (Fig. 7) is not that it is there, something that many would explain with the Balassa-Samuelson effect because of the rapid growth rate of the country. It is that it seems to emerge both in the goods (tradable sector) and services (non-tradable) subindexes, something that initially seems to refute the Balassa-Samuelson argument (although to a certain extent, tourism that constitutes a significant part of services should be considered also as a "tradable service").

An expository comparison with Ireland, where the inflation of goods is much lower than the inflation of services and that thus emerges as a textbook Balassa-Samuelson case, is most revealing.

The high inflation of Greece therefore seems to emerge as a result more of the demand increase, which is largely driven by the expansion of credit and the inflows from the EU structural funds as well as from tourism and shipping industry or



Efficient Market, Fig. 8 Services balance as % of GDP. Eurostat. Goods balance as % of GDP. Eurostat

public borrowing, which is not matched by a similar increase in the domestic supply of goods and services. And this is unlike the case of Ireland in which the surplus of the goods balance seems to finance a deficit in the services balance following again a pattern that fits well the standard predictions of the Balassa-Samuelson model.

The second piece of evidence that supports this argument is the excessive – and increasing – deficit of the goods trade balance, as a percentage of GDP (Fig. 8).

As a matter of fact, the deficit is of a magnitude that has never been seen in any country without the subsequent emergence of serious consequences. In the case of Greece, participation to the eurozone seems to have averted developments like the entrance into a spiral of high inflation and currency devaluations. As a result, the trade deficit in Greece can clearly demonstrate the existence of a serious discrepancy between the growth of domestic demand and the increase of the domestic supply of both goods and services. It should be stressed that in the case of non-tradable services, the inflation differential is sufficient to document the discrepancy between supply and demand, but the emergence of such a differential for goods as well suggests the peculiarity of the case of Greece. Therefore, the evidence at hand would make it more appropriate to label Greece as a unique case of "quasi Balassa-Samuelson," where exports are replaced by EU transfers and domestic credit expansion and the price level is pushed upwards both in the goods and in the services sector, which would actually be in line with the conclusions of recent research on the topic (Gibson 2007; Pelagidis and Toay 2007). The increase of the goods deficit follows as a natural consequence in this case, as increases in demand are satisfied by competitive and

available imported goods as there is no sufficient domestic supply of goods that can compete with the imports.

The third piece of evidence is the following: This persistent deterioration of the goods balance has been financed, besides from the surplus of the services account through foreign inflows in both Greek government bonds and into the stocks of Greek companies, at least until the present financial turmoil. However, it should be noted, rarely these inflows were FDIs. FDIs during the last 3 years were close to zero ($0.9 bil. for 2006, $2.5 bil. for 2007, and $1.3 bil. for 2008, according to the Bank of Greece).

In fact, FDI inward flows for Greece as a percentage of GDP are very low for almost all years, something that is in line with the link between the attractiveness of the business environment and FDI (as described by authors such as Hajkova et al. 2007).

The performance of the goods balance together with the inflation differential with the eurozone for tradable goods suggests also that the cost of importing and distributing these competitive imported goods is higher compared to the eurozone. Furthermore, it suggests that the imports remain competitive in the domestic market in spite of this high cost of importing and distributing, which seems to be really damning for the competitiveness of the domestic supply of goods.

In spite of the mitigating effect of the surplus of the services balance, which is mainly driven by the performance of the shipping industry and tourism, the current account balance has remained for many years at a level (15% to GDP) that in any other country would have been associated with serious consequences. For the two sectors that contribute to the services account surplus, it should be noted that they are less affected by the regulatory environment of the Greek economy, either because they operate almost completely outside the Greek jurisdiction and administrative reality, for the case of shipping, or because they draw their competitive strength largely from the geographical attractiveness and the cultural heritage of Greece, as is the case for tourism.

These pieces of evidence manifest themselves in the compelling case for the low competitiveness of the Greek economy that is documented by a number of surveys (Fig. 9). The impressive part to note here is that a wide selection of different surveys, including those that measure governance and corruption, rank Greece in a roughly similar way even though they often use different methods that are either based on the evaluation of hard evidence, the responses to questionnaires, or a combination of both (Fig. 9).

Facets and Evidence of Institutional Inefficiency and Poor Governance

The OECD Regulation Database, the World Economic Forum competitiveness survey, the World Bank "Doing Business" and Governance Indicators, and the European Commission estimates (EU 2002, 2006), to name a few, all find that in Greece the administrative burden is also exceptionally high (Fig. 10), that regulation of markets is excessive, that government intervention limits competition as well as resource allocation and pricing decisions in crucial network industries, and that the regulation of professional and legal services (Figs. 11 and 12) is high as far as entry and price setting is concerned. At the same time, qualitative standards are excessively lax (Paterson et al. 2003) and that the business environment, as an aggregate, is unattractive.

These findings are complemented by more general statements that indicate weak institutions, poor governance (Kaufmann et al. 2005), and high levels of corruption that seem to follow as a consequence of the high administrative burden and the poor governance (Ackerman 2006).

As a matter of fact, the magnitude of the weaknesses documented by these pieces of evidence matches the size of the competitiveness deficit documented for Greece by the inflation differential with the eurozone, the current account deficit, and the low level of FDIs. It has to be added that, not surprisingly, Greece is found to be the OECD country which has the most to gain from rectifying these documented deficiencies, like product market regulation (Conway et al. 2006), in terms of

Doing Business in 2009, World Bank.	Ease of Doing Business Rank.	World Economic Forum 2008	GCI 2008-2009 rank	Transparency International	2008 Corruption Perceptions Index Country Rank	UN	Rank per capita income in US $
No of countries	181		134		180		214
Greece /total	53%		50%		32%		19%
Singapore	1	United States	1	Denmark	1	Luxembourg	2
New Zealand	2	Switzerland	2	New Zealand	2	Norway	4
United States	3	Denmark	3	Sweden	3	Iceland	6
Hong Kong	4	Sweden	4	Singapore	4	Ireland	7
Denmark	5	Singapore	5	Finland	5	Denmark	8
UK	6	Finland	6	Switzerland	6	Switzerland	10
Ireland	7	Germany	7	Iceland	7	Sweden	13
Canada	8	Netherlands	8	Netherlands	8	Netherlands	14
Australia	9	Japan	9	Australia	9	Finland	15
Norway	10	Canada	10	Canada	10	Australia	16
Iceland	11	UK	12	Luxembourg	11	UK	17
Japan	12	Austria	14	Austria	12	United States	18
Sweden	17	Norway	15	Hong Kong	13	Austria	19
Belgium	19	France	16	Germany	14	Belgium	22
Switzerland	21	Taiwan	17	Norway	15	Canada	23
Estonia	22	Australia	18	Ireland	16	Australia & NZ	24
Korea	23	Belgium	19	UK	17	Germany	25
Mauritius	24	Iceland	20	Belgium	18	France	26
Germany	25	Ireland	22	Japan	19	Italy	32
Netherlands	26	New Zealand	24	USA	20	Japan	33
Austria	27	Luxembourg	25	Chile	23	Spain	35
France	31	Chile	28	France	24	New Zealand	37
South Africa	32	Spain	29	Uruguay	25	Hong Kong	39
Slovakia	36	China	30	Slovenia	26	**Greece**	**40**
Chile	40	Estonia	32	Spain	30	Cyprus	42
Hungary	41	Czech Rp	33	Cyprus	31	Bahrain	43
Tonga	43	Thailand	34	Portugal	32	Puerto Rico	45
Armenia	44	Kuwait	35	Dominica	33	Israel	46
Bulgaria	45	Tunisia	36	Taiwan	39	Slovenia	47
United Arab Emirates	46	Cyprus	40	South Korea	40	Portugal	48
Romania	47	Puerto Rico	41	Latvia	52	Czech Republic	55
Portugal	48	Slovenia	42	Slovakia	53	Estonia	56
Spain	49	Portugal	43	South Africa	54	Saudi Arabia	60
Luxembourg	50	Lithuania	44	Italy	55	Hungary	61
Turkey	59	Slovak Rpb	46	Seychelles	56	Slovakia	62
Italy	65	Italy	49	**Greece**	**57**	Antigua	63
Dominica	74	Turkey	63	Lithuania	58	Latvia	66
Albania	86	Brazil	64	Poland	59	Lithuania	67
Marshall Islands	93	Montenegro	65	Turkey	60	Croatia	68
Serbia	94	Kazakhstan	66	Namibia	61	Poland	69
Papua New Guinea	95	**Greece**	**67**	……		Russian Federation	73
Greece	**96**	Romania	68	……		Venezuela	74
Dominican Republic	97	…….		Sudan	175	….	
……		…….		Afghanistan	176	…..	
…….		Mauritania	131	Haiti	177	Liberia	211
Guinea-Bissau	179	Burundi	132	Iraq	178	Zimbabwe	212
Central African Republic	180	Zimbabwe	133	Myanmar	179	Congo	213
Congo, Dem. Rep.	181	Chad	134	Somalia	180	Burundi	214

Efficient Market, Fig. 9 Competitiveness indexes

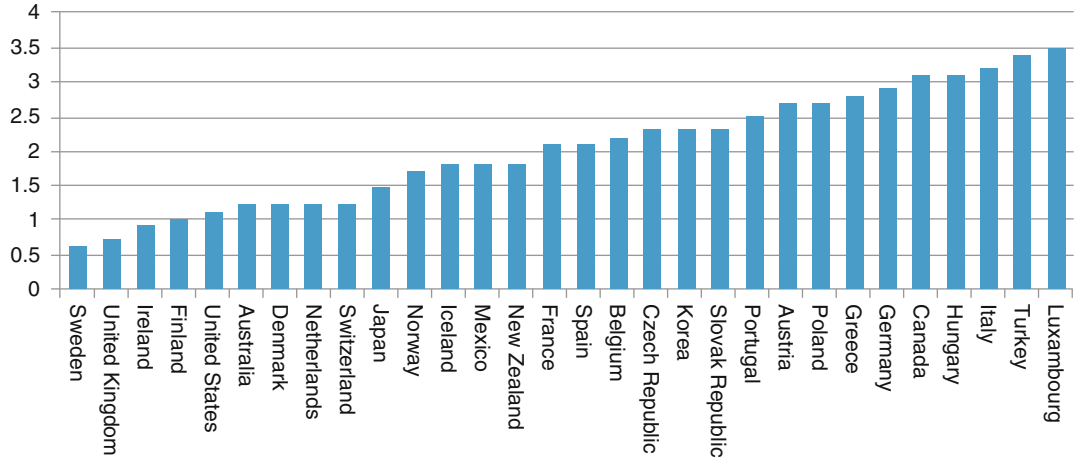
increased productivity. This performance can be labeled "dismal" not because of its absolute level, but because of the large discrepancy between the performance of the country on all these aspects and the per capita GDP that it has achieved in the past years. In particular, following the strong performance till the 1970s and the strong performance of the past years, per capita GDP is

	AT	BL[2]	CZ	DE	DK	ES	FI	FR	UK	GR	HU	IE	IT	NL	PL	PT	RE[2]	SK	SI	SE	EU-25
Administrative cost share in GDP (in%)[1]	4.6	2.8	3.3	3.7	1.9	4.6	1.5	3.7	1.5	6.8	6.8	2.4	4.6	3.7	5.0	4.6	6.8	4.6	4.1	1.5	3.5

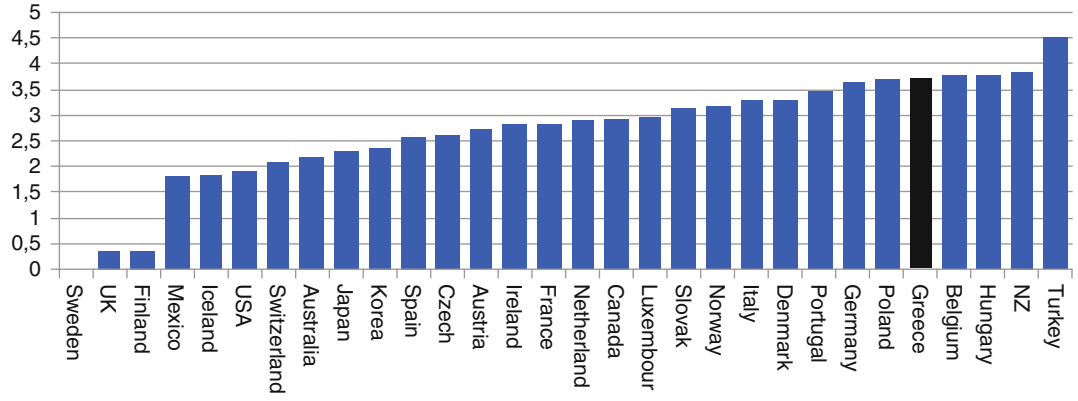
[1] Based on Kox (2005): Intra-EU differences in regulation-caused administrative burden of companies. CPB Memorandum 136. CPB,The Hague.

[2] BL combines Belgium and Luxembourg; RE combines the Baltic Member States, Malta and cyprus; EU-25 figures are GDP-weighted averages

Efficient Market, Fig. 10 Administrative costs by Member State



Efficient Market, Fig. 11 Regulation in professional services (Source: OECD indicator for regulation in professional services, 2007)



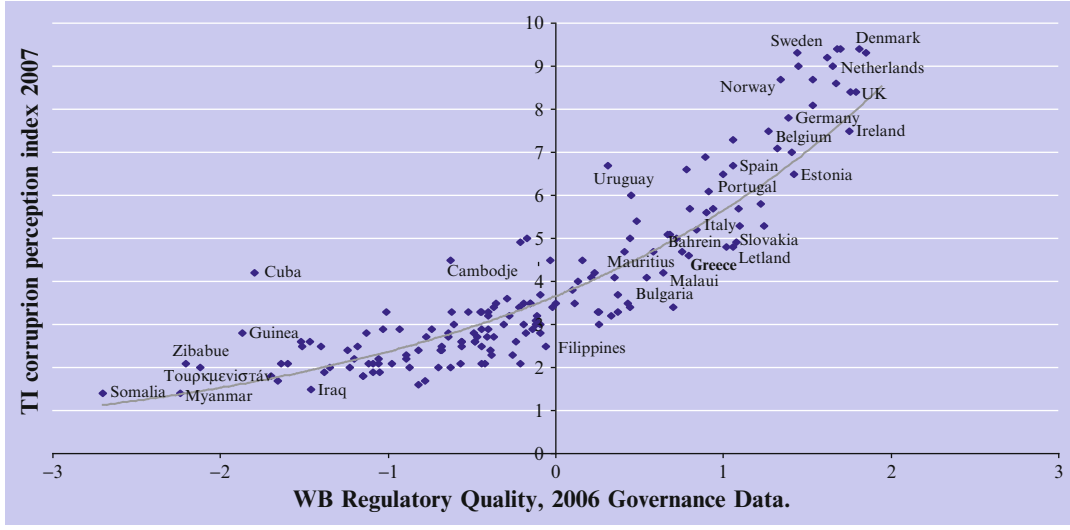
Efficient Market, Fig. 12 OECD indicator. Regulation of legal services, 2007 (Source: OECD. 6 = most stringent)

relatively close to the per capita GDP of the other OECD and EU member countries. And while Greece remains among the poorer members of these groups, it still can distance itself clearly from most other countries that do not participate in these two groups of privileged countries. On the other hand all the other performance indicators are clearly much weaker than the performance of all other OECD and EU member countries. Here Greece clearly is placed, repeatedly, in the middle of the sample of all the countries in the world, and not in the top 20% of the countries, as is the case with per capita GDP. Greece, ultimately, emerges as a country with almost first-class per capita GDP – until at least 2009 – but clearly second-class governance, institutions, business environment, and corruption (Fig. 13).

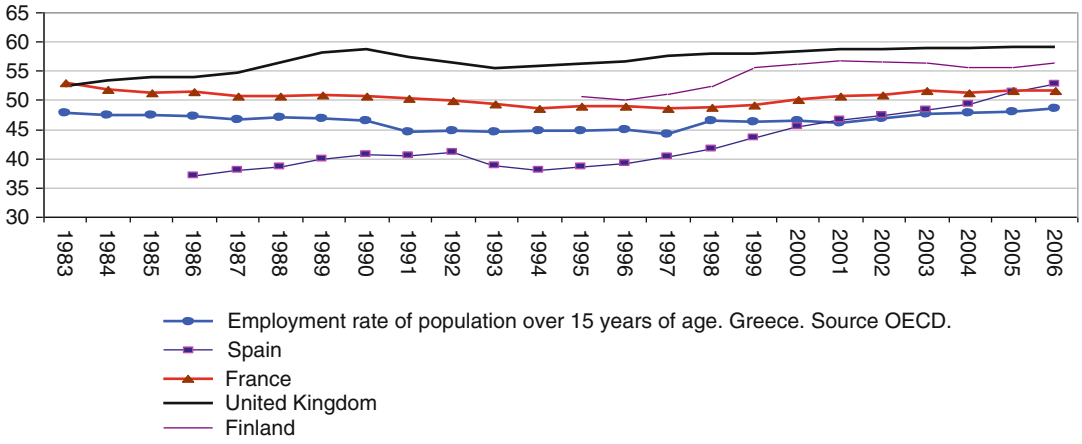
The factors that were analyzed previously and that document why Greece grew so fast in spite of these shortcomings can also reconcile the recent performance of Greece with the now extended literature, mainly of OECD Economic Department Working Papers (an indicative selection of related OECD and non-OECD related publications is OECD (2007a), Conway et al. (2006), Bassanini and Duval (2006), Nicoletti and Scarpetta (2005, 2006), Conway et al. (2005), Bassanini and Ernst (2002), Scarpetta and Tressel (2002), Scarpetta et al. (2002), Nicoletti and Scarpetta (2003), OECD (2003), Nicoletti et al. (2001), Conway et al. (2006)), which directly link the performance of an economy with the quality of the regulatory framework and the prevalence of competitive markets. In a similar way one can reconcile also almost all of the other weak performances of the country that range from research and innovation (Bassanini et al. 2000) to the protection of the environment, the quality of public health services and schools, and the performance of the higher education system (Bassanini and Scarpetta 2001; Mitsopoulos and Pelagidis 2007a; OECD 2007b). Even the weak performance of the judiciary which we will present below in section "Last but Not Least: Inefficiency of Justice" can be ultimately linked to this pattern (Mitsopoulos and Pelagidis 2007b; Djankov et al. 2002).

The Paradox of the Underlying "High Labor Productivity" in a Low Competitiveness and Inefficiency Context

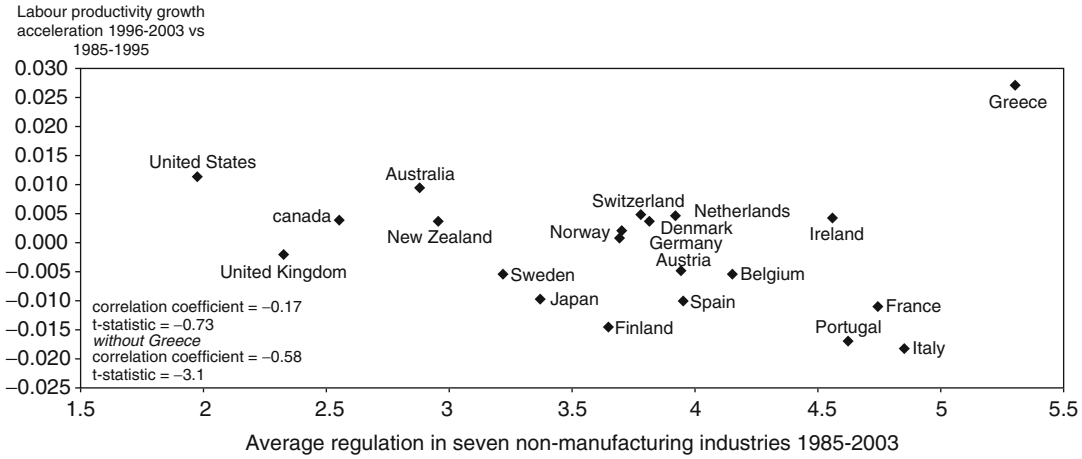
The result of the strong demand growth that is not driven by an increase in domestic supply that follows from an increase in employment (Fig. 14) directly affects the reliability of



Efficient Market, Fig. 13 Corruption and Regulation



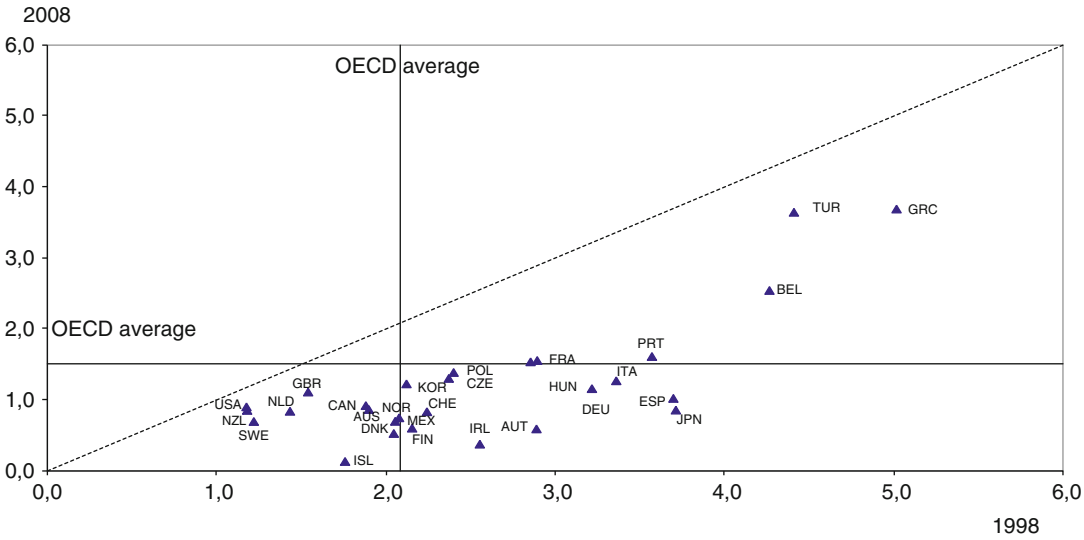
Efficient Market, Fig. 14 Employment ratio for the population over 15 years of age



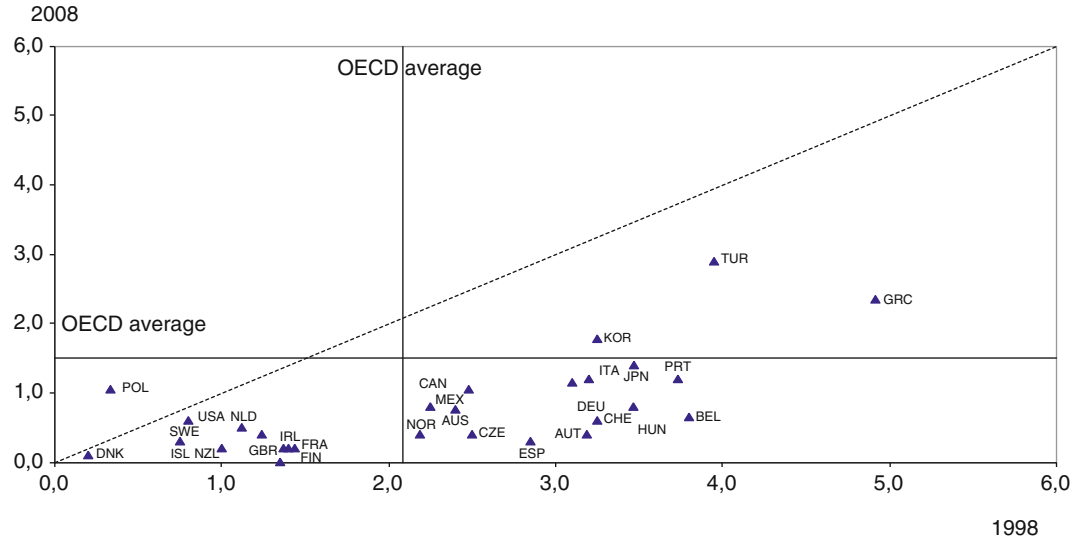
Source: OECD Productivity Database and OECD International regulation database

Efficient Market, Fig. 15 The scale of the indicators is 0-6 from least to most restrictive

productivity indexes that measure GDP to labor input – and that give a% of around 2.5–3% for Greece during these years – in various forms. This follows as the increase in the numerator (GDP) matches a restrained increase in the denominator (as can be seen in Fig. 14), thus measuring a large increase in the productivity per worker or per hour worked, in spite of the dismal performance of the Greek economy as measured by the rigidity index of relevant product markets (Fig. 15). It follows from the previous exposition that the use of such indicators is not correctly capturing the variety of the parameters that shape the performance of the Greek economy during the past decade, often depicting Greece in a position that does not favor the drawing of reliable conclusions. This gives also an explanation to the puzzle of having on the one side high GDP and productivity rates and on the other side low competitiveness with twin deficits. At least to the extent, we take into account only domestic forces and not taking into account factors such as euro's overvaluation (at least to the extent that Greece's trade takes place with outside EU partners (around 50% of total)) and the asymmetric demand shocks.



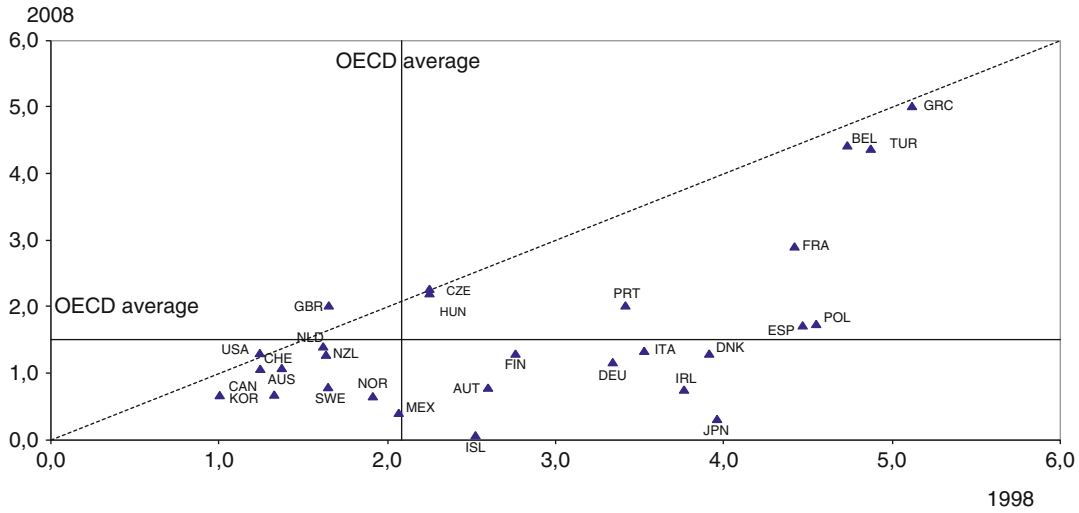
Efficient Market, Fig. 16 State control. Involvement in business operation (Source: OECD. Index scale of 0–6 from least to most restrictive)



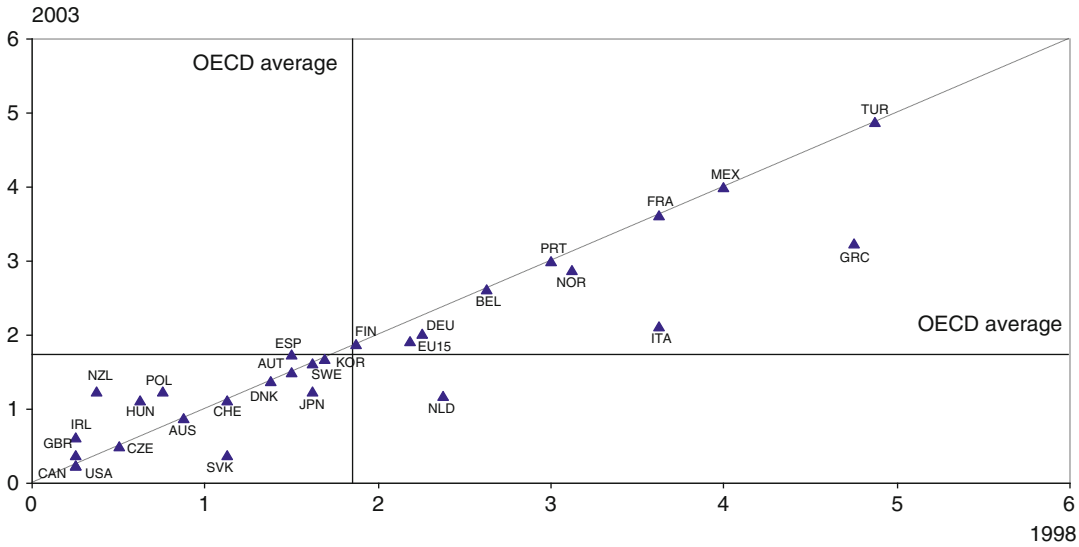
Efficient Market, Fig. 17 State control – price controls (Source: OECD. Index scale of 0–6 from least to most restrictive)

In this context, it is worth looking more on some other aspects of institutional rigidities which complement very well low competitiveness. Figure 16 summarizes product market regulation regarding state control through the involvement in business operation. Once again Greece appears to both have the most stringent state involvement in the operation of businesses and to maintain these practices throughout the past decade.

Figures 17 and 18 concern the state involvement in business operations via price controls or the use of command and control regulation. "Command and control" includes a lot of



Efficient Market, Fig. 18 State control. Use of command and control regulation (Source: OECD. Index scale of 0–6 from least to most restrictive)



Efficient Market, Fig. 19 Employment protection legislation (However, the above data does not include the so-called shadow economy, which in the case of Greece it is estimated around 20–30 % of the official GDP. The "black labour market," as part of the shadow economy, includes a significantly flexible labour market that excludes employees from trade unions' membership, social protection and insurance, working rights, etc.)

administrative mechanisms of hindrance of entrepreneurial activity/organization, in sectors such as "road and railway transports" and retail trade.

Product market rigidities are of critical importance for rigidities in the labor markets as well. Figure 19 shows Greece among OECD countries with the highest employment protection legislation (EPL). It should be noted here that the market for nonpermanent, temporary employment in Greece is the main reason for the exceptional rigidity of the Greek labor market overall but that the market for permanent contracts is also

relatively rigid when compared with other OECD countries.

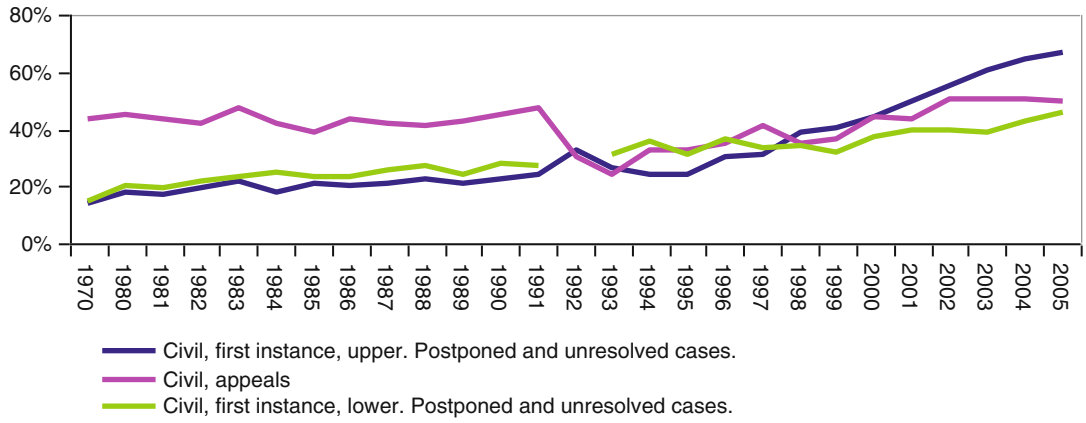
These kinds of structural institutional rigidities/inefficiencies constituted a true cost to society in the environment of a noncompetitive economy like the Greek economy. It meant and led to the exclusion of many others from the labor market, and especially the young that seek salaried labor. Under 26 years old unemployment was more than 35% and 20% for women and men correspondingly even before the crisis. This should be read as underutilization of a dynamic labor force and should not be considered solely as a major social or ethical issue.

We mentioned extensive regulation of markets, high administrative costs, a business environment that is not favorable to entrepreneurship and, in the end, weak convergence and widespread corruption as drivers and cause of this low competitiveness and low efficiency, in spite of the reforms in the credit and telecommunications markets and the benefits accruing from EMU accession. Greece emerges, therefore, to benefit from certain reforms in terms of potential output because of the nature and importance of these positive developments, while it retains other, also significant in importance and magnitude, weaknesses that undermine the long-term growth potential of the country. These weaknesses are ultimately described as "rigidities and inefficiencies, weak non-independent institutions and governance," and their proliferation was deeply built in the equilibrium that was formed between the interest groups that accrued the rents that they secured through the regulation of markets and the inflation of the administrative costs (Pelagidis and Mitsopoulos 2009). One could also argue that the strong growth of the past years has also made the need for further reforms less pressing.

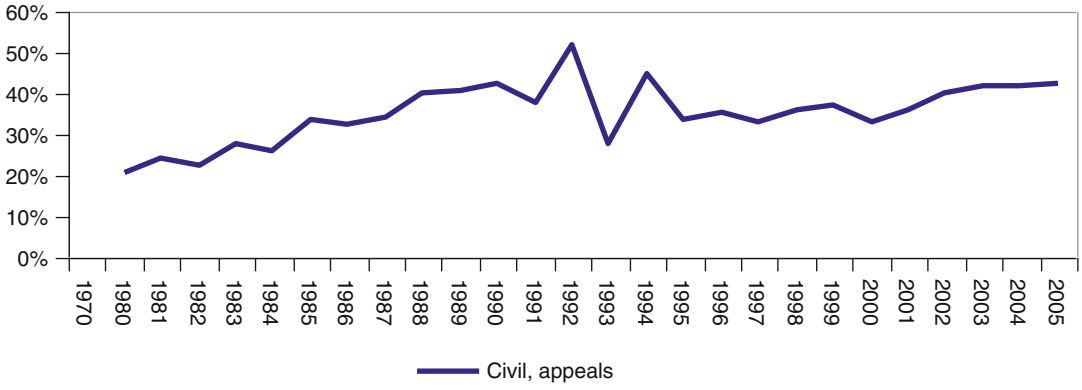
Last but Not Least: Inefficiency of Justice

The weak performance of the judiciary, which we present very briefly below, can be ultimately linked also to this pattern (Mitsopoulos and Pelagidis 2007b; Djankov et al. 2002).

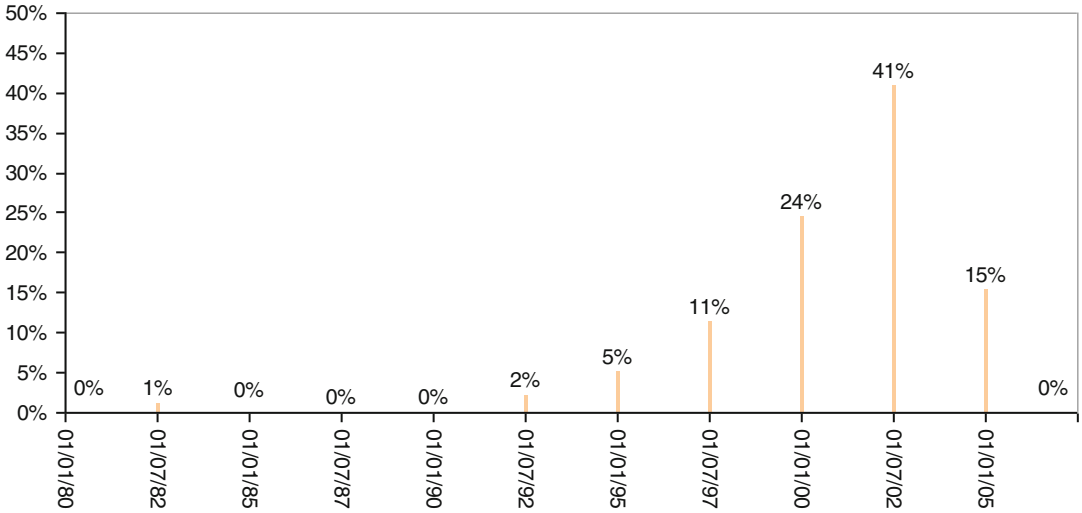
This unfortunate picture refers to cases that are quite old; therefore, the deterioration in the past years of the indirect measures of the time needed to dispose justice used in Mitsopoulos and Pelagidis (2007b) suggests that the cases that now enter the system may face even worse prospects. In particular, Fig. 20 shows that the ratio of cases that have not been heard to the sum of cases that are either pending or were introduced for the first time in the year has increased in upper civil courts from about 25% in 1997 to over 60% in 2005. The deterioration is visible, but less dramatic, in appeals courts and lower civil courts. Also, as shown in Fig. 21, for civil courts the appeal rate has increased slightly from 1997 to 2005.



Efficient Market, Fig. 20 Unresolved and postponed cases to total cases introduced



Efficient Market, Fig. 21 New introduced to appeals courts to concluded by first instance courts



Efficient Market, Fig. 22 % of cases first filed from the latter data of the range and till the latest date of the range. Cases to be retried

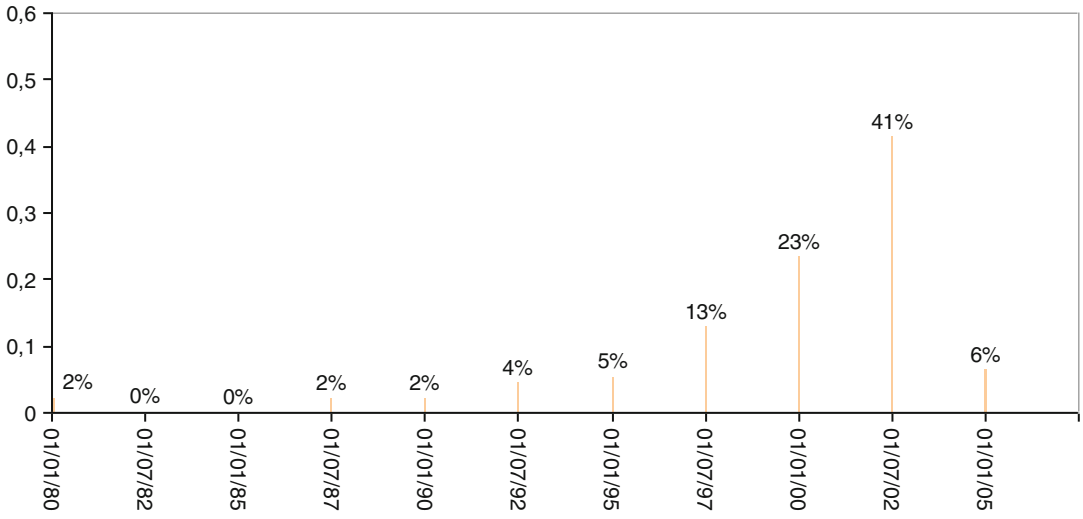
We also note in our sample in Mitsopoulos and Pelagidis (2010) for the Greek Supreme Court (*Areios Pagos*) some extreme delays in the process of serving justice (Fig. 22). Twenty percent of our cases have not been settled after almost 10 years since they were first filed, and even for these cases the process will be delayed for at least another year till the appeals court issues a new verdict. Sixty-five percent of the cases await a new verdict from the appeals court in spite of the fact that they were first filed between 5 and 10 years ago. On average, at the end of the year 2006, the cases tried by the *Areios Pagos*, and that were marked for a retrial, were 6.5 years old. We also find a small but not insignificant number of cases, 8% of our sample, in which we document repeated appeals to the civil *Supreme Court*. These originate among the 40% of all the cases that have been retried by the order of the *Areios Pagos*. It seems that for a number of these, the new decision of the appeals court, or the court that tries in its place, is challenged anew by one of the involved parties in front of the *Areios Pagos*. While the total number of these cases is small, about 8% of the total, they are about half of the cases that were initiated more than 12 years before 2006 and about one fourth of

the cases that are over 10 years old. This indicated that there is a small, but non-negligible, chance that a case will get entangled in a process of repeated retrials, and this process ensures that no final verdict is obtained for at least a decade. Our sample contained two cases that were judged for the third time by the *Areios Pagos* and thus are due to get tried for the fourth time by the appeals court, and one case where the second verdict of the *Areios Pagos* got challenged and was directly brought again to the court for retrial.

Such delays also apply to cases in which the Supreme Court reaffirms the decision of the previous court. Taking a random sample of 100 of these cases from the year 2006, we see that these cases, which form about 60% of the cases in the year 2006, were on average over 7.5 years old at the time the decision of the *Areios Pagos* was published. And while 70% of these cases was less than 7 years old, a significant 30% was older than 7 years with one case being first introduced almost 26 years ago and another case 33 years ago (Fig. 23). About 25% of these very old cases, which were first introduced more than 7 years ago, were cases that had been retried by the appeals court before, after a previous order to do so, but 75% of these old cases simply proceeded very slowly.

These long delays in the process of official arbitration are not caused because one of the involved parties is not satisfied with the outcome, say because the height of the compensation awarded did not meet its expectations. They are caused because the lower courts either do not follow procedures properly or because they ignore, or misinterpret, laws which is an objective measure for the low quality of the decisions that are taken by the lower courts. In itself though, the use of the appeals process for error correction, as opposed to first instance and appeals courts decisions that do not make mistakes, does not provide by itself information on the quality of the judicial system, as explained by Shavell (1993).

Measuring the quality is curtailed by the fact that we have a complete sample only from *Supreme Court* decisions that do judge only the lawfulness of the preceding court decisions, as is appropriate for a civil law country. Also, since we are not always informed about the full content of all the preceding court decisions, that is, both the appeals and first-level court, or in the case of previous retrials the content of the initial court decisions, we cannot construct measures that compare the initial with the subsequent decisions. Furthermore, the structure of civil law does not permit judges to commit errors that may be



Efficient Market, Fig. 23 % of cases first filed from the latter data of the range and till the latest date of the range. Cases reaffirmed

brought before the highest court that we may use to pinpoint a low quality of court decisions that refer to issues like the height of compensation. Therefore, we are left once again only with a measure of time that allows us to measure the time needed to serve finally justice. As in our previous work (Mitsopoulos and Pelagidis 2007b), where we employed an indirect measure for time, we find again, using this time a direct measure for the time needed to dispose justice, that in Greece justice is served often with delays that are excessive and which undermine the usefulness of the official arbitration mechanism.

The 2010–2013 "Troika Period": Reforms for Achieving Efficiency

Below we offer a list of reforms already implemented by the Greek government in the context of the conditionality programs agreed with the troika and the creditors of the country during the period 2010–2013. The list is not exhaustive albeit representative. Reforms are leading the country towards a much more efficient economic structure.

Labor Market Reforms

- Significant reforms for the private sector from 2010 (reduction of severance payment, relaxing of layoff limits, activation of tools to reduce individual salaries, possibility to bypass general wage agreements at the firm level) have increased downward wage cost flexibility.
- Even more significant reforms tied down in MoU2 (rationalization of mediation process, reduction of minimum wage, and impact of sectorial and other agreements).
- More fine-tuning reforms rumored for 2013–2014.
- Much lower minimum wage.
- Large reduction in firing costs for high earners.
- Redefining threshold for group layoffs.
- Removal of all impediments to conclude wage agreements at the company level.
- Significant changes in the mediation process.
- A very important impact of these in the reality behind average labor cost indexes.

In the end, these reforms do establish a truly flexible and efficient labor market in Greece.

Fiscal Reforms

- Wide ranging social security reform
- Large cuts in public sector salaries and salaries of public utility companies
- Healthcare reform (online prescriptions, pricing of medicine, etc.)
- Reforms in budgeting and fiscal reporting
- Very large, across the board, tax increases (albeit with a significant inflationary impact)
- Accelerating drive against tax evasion

Structural Reforms (selectively)

- Licensing process and spatial planning reform (about 40% complete)
- Road haulage deregulation (about 50% complete)
- Cruise ship home porting regardless of flag
- Already some deregulation in professional fees and obligatory purchase of services (notary public, lawyers for small real estate transactions, company start-ups, fees of engineers, trademark submission without lawyer attendance, sale of baby milk outside pharmacies, etc.)
- Establishment of one-stop shops for company creation and gradual improvement in their operation (yet complicated process design does not work well)
- Export facilitation (abolishing obligation to register in "exporters registry," a strategy to boost exports, link Piraeus port terminal operated by COSCO with rail network, etc.)

An Indicative List of New Taxation Since the First Memorandum 2010

- Multiple abolition of tax exemptions
- Increased of annual estate tax, up to 2% of administratively set value, and then additional estate tax through electricity bills
- Multiple VAT increases
- New personal income taxation law with higher rates, lower tax-free income
- 3 tax increases on tobacco and alcoholic beverages and VAT increase on nonalcoholic beverages

- Further increase of tax on mobile communications – highest in the EU
- 3 gas/petrol tax increases
- Multiple increases of taxes on electricity, especially for businesses – now probably highest energy prices in the EU for industry
- Introduction and increase of luxury tax
- Recurrent extraordinary tax on profitable companies and increased tax on dividends
- Introduction and increase of recurrent extraordinary tax on high personal incomes
- Tax on banks and increase of tax advancement
- Tax on TV advertisements
- Tax on violations of building permits
- Green tax
- Income tax on leased cars and private sector company cars (but not government official cars!!!)
- Introduction and then increase of tax on assumed income
- Increase in train and bus tickets
- Tax to avoid antismoking ban
- Fee to access hospitals
- Highest recurrent real estate taxes in the OECD

Since the initiation of the conditionality programs in 2010, and after two revisions and a debt haircut to private creditors in 2012, the improvement of the business environment, the remove of red tape, and the creation of a level playing field through the removal of restrictive to competition laws have been rather a second priority both for the Greek governments and for the official lenders of the former, when compared with tax increases and labor market reforms and at least for the two first years of the program implementation. While this has been changing slowly during 2013–2014, with the official lenders now demanding from the Greek government progress on these issues and while they now offer also extensive technical assistance, it must be noted that the delay in the acknowledgment of the importance of such measures may yet prove to be crucial. In sum, more reforms are needed in the domains of product and professional markets to increase thoroughly efficiency. Below one can find a few suggestions on this issue.

More Reforms for Efficiency

- Complete licensing, business park legislation, etc.
- Complete professional services deregulation
- Rationalizing the complex system of third-party payments
- Rationalizing regulation in the energy sector
- Rolling out the privatization program in a more aggressive way, emphasizing assets crucial for competitiveness like ports and not gambling
- Structural reforms that will improve the business environment and that are yet to be identified
- Emphasis at last in deregulating product and professional markets
- Last (but not least), a (meaningful) tax reform which will not include today's overtaxation

References

Ackerman SA (ed) (2006) International handbook on the economics of corruption. Yale University Press, New Haven

Bassanini A, Duval R (2006) Employment patterns in OECD countries. Reassessing the role of policies and institutions, vol 486, ECO WP. OCED, Paris

Bassanini A, Ernst E (2002) Labour market institutions, product market regulation and innovation. Cross country evidence, vol 316, ECO WP. OCED, Paris

Bassanini A, Scarpetta S (2001) Does human capital matter for growth in OECD countries? vol 282, ECO WP. OCED, Paris

Bassanini A, Scarpetta S, Visco I (2000) Knowledge, technology and economic growth: recent evidence from OECD countries, vol 259, OECD ECO WP. OECD, Paris

Conway P, Janod V, Nicoletti G (2005) Product market regulation in OECD countries: 1998 to 2003, vol 419, OCED ECO WP. OECD, Paris

Conway P, de Rosa D, Nicoletti G, Steiner F (2006) Regulation, competition and productivity convergence, vol 509, OECD ECO WP. OECD, Paris

Djankov S, La Porta R, Lopez-de-Silanes F, Shleifer A (2002) The practice of justice. World Bank Development Report 2002. Washington, D.C.

EC (2006) Measuring administrative costs and reducing administrative burdens in the European Union. Commission Working Document COM (2006) 691 final, 14.11.2006, london

EU (2002) Benchmarking the administration of business start-ups. European Commission Final Report, Brussels

Gibson HD (2007) The contribution of sectoral productivity differentials to inflation in Greece. Bank of Greece, WP, vol 63
Hajkova D, Nicoletti G, Vartia L, Yoo K-Y (2007) Taxation, business environment and FDI location in OECD countries. OECD ECO WP, vol 501. Also published in OECD economic studies no. 43/1 2007
Kaufmann D, Kray A, Mastruzzi M (2005) Governance matters IV. The World Bank, Washington, DC
Mitsopoulos M, Pelagidis T (2007a) Rent seeking and ex-post acceptance of reforms in higher education. J Econ Policy Reform 10(3):219–44
Mitsopoulos M, Pelagidis T (2007b) Does staffing affect the time to serve justice in Greek courts? Int Rev Law Econ 177–192
Mitsopoulos M, Pelagidis T (2010) Greek Appeals Courts' Quality Analysis and Performance. Eur J Law Econ 30 (1):17–39
Nicoletti G, Scarpetta S (2003) Regulation, productivity and growth. OECD evidence, vol 347, OECD ECO WP. OECD, Paris
Nicoletti G, Scarpetta S (2005) Product market reforms and employment in OECD countries, vol 472, OECD ECO WP. OECD, Paris
Nicoletti G, Scarpetta S (2006) Regulation and economic performance: product market reforms and productivity in the OECD, vol 460, OECD ECO WP. OECD, Paris
Nicoletti G, Bassanini A, Ernst E, Jean S, Santiago P, Swaim P (2001) Product and labour markets interactions in OECD countries, vol 312, OECD ECO WP. OECD, Paris
OECD (2003) The sources of economic growth in OECD countries. OECD, Paris
OECD (2007a) Going for growth. OECD, Paris
OECD (2007b) Economic surveys: Greece. OECD, Paris
Paterson I, Fink M, Ogus A (2003) Economic impact of regulation in the field of liberal professions in different member states, regulation of professional services, final report-part3, Jan 2003, study by the Institut fuer Hohere Studien, Wien for the European Commission, DG Competition
Pelagidis T, Mitsopoulos M (2009) Vikings in Greece. Kleptocratic interest groups in a rent-seeking society. Cato J 29(3):399–416
Pelagidis T, Toay T (2007) Expensive living: the Greek experience under the euro. Intereconomics Rev Eur Econ Policy 42(3):167–176
Scarpetta S, Tressel T (2002) Productivity and convergence in a panel of OECD industries. Do regulations and institutions matter? vol 342, OECD ECO WP. OECD, Paris
Scarpetta S, Hemmings P, Tressel T, Woo J (2002) The role of policy and institutions for productivity and firm dynamics: evidence from micro and industry data, vol 329, OECD ECO WP. OECD, Paris
Shavell S (1993) The appeals process as a means of error correction. J Leg Stud 12(2):379–426

Eighteenth-Century Piracy

► Piracy, Old Maritime

Electoral Systems

Guido Ortona
Dipartimento di Giurisprudenza e Scienze Politiche, Economiche e Sociali, Università del Piemonte Orientale, Alessandria, Italy

Definition

Electoral system. An electoral system is the set of rules employed to summon a representative body on the basis of the preferences expressed by a group of electors, through a given voting procedure. The chapter compares electoral systems with reference to the election of a Chamber of a democratic Parliament.

Introduction and Overview

An *electoral system* is the set of rules employed to summon a representative body on the basis of the preferences expressed by a group of electors, through a given voting procedure. This definition applies to many settings, from parents' circles to supernational organizations; in this survey we will deal only with the election of a Chamber of a Parliament. We will also limit the discussion to the context that is usual in present-time democracies; hence we will assume (a) that all voters express the same number of votes (very often one, but there are many exceptions) and (b) that within the set of allowed alternatives, there are no constraints on the choice made by the voters. The votes are cast in an *election,* and for obvious reasons the election usually takes place in different *districts;* there are however several countries, not all of them minuscule, where there is a unique nationwide district. This is the case, for instance,

of Israel and the Netherlands. The number of the Members of Parliament (normally shortened to MPs) to be elected in each district is the *district magnitude*.

A meaningful classification of electoral systems may be obtained through the crossing of two dual dimensions, *nominality* and *plurality/proportionality.* A district may elect one MP, in which case it is *uninominal*, or more, and it is *plurinominal*; and the Seats may be assigned either *proportionally* to the valid votes obtained by each list of candidates (usually a *party*) or according to the number of votes received by each candidate, in which case the system is *majoritarian*. A uninominal election is obviously majoritarian, while a plurinominal one needs not being proportional but usually is, hence the frequent identification in the everyday political debate of uninominal elections with the use of the majority rule and of plurinominal ones with proportionality.

The electoral law plays a fundamental role in the functioning of a democracy, and it is a main determinant of the perspectives, and hence of the strategies, of both incumbent MPs and challenging candidates. No surprise, then, that the electoral law may be subject to changes that may easily be dramatic and frequent. Comprehensibly, this is particularly true in times of political and economic turbulence; this entry is being written in such a period, hence the reader is reminded that the data quoted in it are those of the last months of 2016.

An authoritative source (IDEA 2005) proposes the following classification of the electoral systems actually employed in at least one country at the time of its publication:

Plurality/Majority systems: Plurality (also labeled first-past-the-post), two-round system, alternative vote, block vote, and party block vote
Proportional representation: List proportional (open or closed), single transferable vote
Mixed systems: Parallel, mixed-member proportional
Other: Single nontransferable vote, limited vote, and Borda count

The Mixed member proportional system is the system in use in Germany; the author of this entry suggests that it should be considered a member of the proportional representation family. More on this in section "Proportional Systems." The list is not exhaustive, more so if one considers also the systems that have been defined by the doctrine, those employed at a sub-parliamentary level, and the variants within a given system. What is remarkable is the absence, among the plurality systems, of the Condorcet system. The Condorcet system qualifies as the best of the family because if the choice is not the Condorcet winner, then there exists another one that a majority prefers to it. (The *Condorcet winner*, i.e., the alternative chosen when the Condorcet procedure is in use, is the one that is preferred by a majority in dual contests against each of the others in turn. It may not exist, but in large electoral settings this is a very rare occurrence, and it may be taken into account via a suitable tie-breaking rule.) The doctrine assumed the Condorcet system as the benchmark to be employed to assess comparatively other plurality systems, since the very beginning of the quantitative comparison of electoral systems up to now (see Merrill 1984, and Diss and Doghmi 2016). Not surprisingly, normally the best system results being the *Borda rule* (see below). The implementation of the Condorcet rule requires computing facilities not available until relatively recently, yet the increasing introduction of electronic voting devices makes it a reasonable option. No country, however, is presently considering its adoption.

In this entry we will describe the aforementioned systems in some detail (sections "Majoritarian Systems," "Proportional Systems," and "Mixed and Other Systems"), and then we will move to the problem of the choice of the electoral system (sections "Majority or Proportionality? Theory," "Majority or Proportionality? Evidence," and "Electoral Reforms and Public Opinion").

Majority, Proportionality and History

The most relevant difference among electoral systems is the fundamental principle upon which they are based, that is whether they are *majoritarian* or *proportional.* Before proceeding, it is necessary to

make it clear a very basic point, which is not always duly considered in the literature, let alone in the political debate. A majoritarian system assumes that the citizens of a particular constituency have common interests and/or opinions and that they are called to choose a delegate to represent at the best their interests and/or opinions, in the spirit of the *theorem of Condorcet*. (The theorem states that the probability of making the best choice increases with the number of deciding subjects, provided – inter alia – that there is a choice that is actually the best for each of the subjects.) A proportional system assumes instead that the community of interests does not concern a given location but a given *social group* throughout the territory. At the district level, the communality of interests may not exist, and if it exists, it is inferior to other, nonlocal communalities. It follows that the representation must be divided (proportionally) among the various groups, in order to reproduce the articulations of the society in a manageable scale. This is why Parliaments with an ancient origin are typically majoritarian, like those of the UK and of the USA, while modern ones are typically proportional (for a discussion of this point based on historical evidence, see Ahmed 2013). Also, this is why the majority of present-time scholars favor proportionality (see Bowler and others 2005).

Worldwide, in 2016, 85 Parliaments were proportional, 62 were majoritarian, and 34 were mixed (including Germany); the use of other systems was not uncommon, as they summed to 69, divided among nine different systems (data from the ACE Electoral Knowledge Network, http://aceproject.org/). Proportional systems are most largely employed in Europe, where 35 Parliaments (including Germany, see below) out of 46 belonged to the family.

Majoritarian Systems

The simpler and most used majoritarian system is uninominal *plurality*, also labeled *first-past-the-post* (FPTP); it is adopted in the UK, USA, India, and in many other countries of British tradition. There are as many districts as MPs to be elected, and in each district the candidate who obtains a plurality of valid votes is elected. The *two-round system* is adopted in France and in several countries formerly in the French empire. The districts are uninominal, but if no candidate obtains a *majority*, there will be a second round involving a subset of the candidates (in France those who got at least 12.5% of valid votes, or the first two if none exceeded this threshold). The other majoritarian systems are much less in use. With the *alternative vote* (also known as *instant-runoff vote* or *transferable vote*), the districts are uninominal too. The voter is requested to rank all the candidates according to her/his preferences or some of them. In the first round only the first preferences are considered, and a candidate is elected if she/he obtains a majority. If none gets it, the candidate with less votes is excluded, and each of her/his votes is transferred to the second preference of the voter. The procedure is iterated until the majority arises. The system is employed in Australia and in other countries in Oceania; in 2011 a referendum rejected its adoption in the United Kingdom. *Block vote* and its variant *party block vote* are majoritarian but plurinominal. In the first version each voter may vote for several candidates, and those with more votes are elected; in the second each voter votes for a party, and the party with more votes wins all the seats of the district. The first variant is in use in Lebanon and Mauritius, but the most extreme application occurs in Monaco, where 16 MPs out of 24 are elected with this system, and each voter may cast up to 24 votes. The second variant is in use in Cameroon and Chad.

Proportional Systems

A *pure proportional system* (one national district, no seat assignment threshold) is in force only in 10 countries, that is Colombia (Senate only), Fiji, Israel, Mozambique, Namibia, the Netherlands, Paraguay (Senate only), Serbia, Slovakia, and South Africa. In any case, the seats are assigned to each party according to the share of valid votes obtained. With proportionality there is a problem with the rounding of the results, that

may be serious if the district magnitude is small. Most countries employ the *D'Hondt rule,* a member of the family of *highest average methods.* The number of valid votes of each party in each district are divided by successive integers, starting with 1, and the M largest ratios, where M is the district magnitude, individuate the elected candidates. There are several variants; the one most in use is the *Sainte-Laguë rule*, where only odd integers are considered, what favors the smaller parties. *Largest remainder methods* are much less employed: their principle is that the remainders determine the seats to be assigned to each party after that each one obtained the number of seats given by its integer share of votes. The party list may be either *closed*, when the voter may only choose the party, or (partially or totally) *open*, if she/he may also indicate one or more preferences within the list.

The real proportionality depends crucially from the *district magnitude.* Lijphart suggested that the following formula:

$$T = 75/(M + 1)$$

where T is the implicit percent threshold and M is the district magnitude may roughly indicate the implicit threshold that the magnitude introduces. For instance, a district magnitude of 9 would produce (roughly) the same results of a district with a threshold for participating to the assignment of seats of 7.5%.

There is a commonsense wisdom that a trade-off exists between *proportionality* and *governability:* a not-constrained proportional vote would produce a plethora of parties, what would make it difficult the summoning of a stable and effective government. This assumption has been criticized by several authors, both on empirical and theoretical foundations, and actually it does not enjoy a sound theoretical basis (see f.i. Migheli et al. 2014). The reason is that a majoritarian system tends to produce few large parties, usually two, that include interests and opinions that may be highly different if not contrasting, in the logic of the so-called *Law of Duverger*; what may induce a bargaining *within* a party that may be as engaging as that *among* parties in proportionality. (The Law of Duverger is not a law but an empirical regularity, not very often respected in full, according to which plurality produces a two-party system.) Nevertheless most countries adopt some correction to pure proportionality; these corrections are of three types, a *participation threshold*, by far the most common, a *majority prize*, and a *reduced district magnitude.*

Participation thresholds (that establish that a party will not get seats unless its share of valid votes is greater than the threshold) range typically around 3–5%, with Liechtenstein (8%) and Turkey (10%) outlying. As stated above, a threshold is in use in most proportional countries.

The majority prize is present in Greece and Italy (and in san Marino). The party (or the coalition of parties) that obtain a given share of votes obtains additional seats. There are some historical precedents, like the *Law Acerbo* (1924, from the name of its author), that opened the way to the electoral triumph of Mussolini and the so-called *Loi Scélérate* in France and *Law Truffa* in Italy ("scélérate" meaning wicked and "truffa" meaning fraud), respectively, in 1951 and 1953. Both the historical precedents (and the names they were popularly labeled in the last two cases), and the inability to resolve a political stalemate in the two countries concerned at present, justify clearly enough the limited consensus that this system enjoys among the scholars. In Spain the districts are usually small, what, as we saw, reduces the proportionality; again the results in terms of governability are apparently poor.

The *single transferable vote* aims at preserving the vote-for-a-person feature of majoritarian systems while at the same time assigning the seats proportionally. This is why it is highly appreciated by the English and American literature; however, by its very nature, the system requires small districts, what makes its proportionality limited. It is in use only in Australia (Senate), Ireland, and Malta. Omitting minor technicalities, its typical structure is the following. Suppose a district magnitude of M, say, 5, and N valid votes, say, 1,200. Each voter must rank all the candidates according to her/his order of preference. All the candidates that are ranked first by at least 201 voters are

elected; this is because at most 5 candidates may obtain so many first preferences (and in general $1 + M/(N + 1)$ first preferences). Suppose now that a candidate obtains 301 first preferences. 201 are "used" to win her/his seat, while 100 (chosen at random) are transferred (whence the name of the system) to the *second* preference, and the procedure is iterated until 5 MPs are summoned.

Mixed and Other Systems

A *parallel system* is in use when some MPs are elected with a system and the others with another one (or possibly more); the two systems are normally a majoritarian and a proportional one. A parallel system is supposed to guarantee a representation, albeit with a reduced share, to all the parties whose share of votes exceeds the established threshold, while at the same time producing an enhanced governability thanks to the majoritarian principle. Several countries adopt a parallel system, including Russia, Mexico, Hungary, the Philippines, Thailand, Venezuela, Japan, and South Korea. Instead, a *mixed-member proportional system* maintains the proportionality while requesting also a uninominal vote. The most relevant example is Germany; voters vote for a candidate in a uninominal majoritarian district and at the same time for a party in a large proportional one. The seats to be assigned to each party are determined by its share in the proportional vote, but the candidates obtaining a plurality in the uninominal districts are elected anyway. If a party obtains, say, 200 seats in the proportional vote but it wins in 205 uninominal districts, the total number of MPs is increased correspondingly. (In Germany a 5% national threshold is in force, but this threshold is not requested for a party that wins in at least 3 uninominal contests.) As a result, the system is not that different from an open list proportional one. New Zealand and Bolivia also employ it.

The *single nontransferable vote* is the application of the majority rule to plurinominal districts. Each voter has one vote, and the M candidates with more votes are elected, M being the district magnitude. It is in use in some Asian countries, among which Indonesia (Senate). The *limited vote system* is analogous, but the voter may express more than one vote, albeit less than the district magnitude; it is adopted in Spain for the election of the Senate. Finally, the *Borda count*, despite the appealing theoretical properties of the system, is employed only in Slovenia, and for two seats only, and in Nauru. (Each voter ranks the candidates, giving 1 point to the first, 2 to the second, etc.; the points are summed, and the M candidates with less points are elected, M being again the district magnitude.)

Majority or Proportionality? Theory

The political doctrine made a lot of efforts to individuate the advantages and the flaws of the different electoral methods. We will consider only the comparison of the two grand families, that of the proportional systems and that of majoritarian ones. Here is a (non-exhaustive) list. For a more detailed discussion of the points listed below, see IDEA (2005), and the sub-site *Electoral Systems* of the site of ACE (Electoral Knowledge Network) or, obviously, the (enormous) scientific literature.

Majority/Plurality, Advantages

(a) It tends to produce a one-party government, what enhances both stability and governability.
(b) Symmetrically, it produces a strong and realistic opposition.
(c) It encourages the parties to represent broad arrays of interests.
(d) It tends to exclude extremist parties from the Parliament.
(e) It forces the parties to choose credible and respected candidates and allows voters to choose individuals instead of parties.
(f) It opens the way to independent candidates.
(g) It is simple to understand.
(h) It enhances the accountability of the government, as the responsibility for unpopular choices cannot be easily concealed.
(i) It promotes the representation of the local interests.

Majority/Plurality, Flaws

(a) Many voters will not be represented.
(b) It may assign all the power to a party that may well have a limited consensus, what could create serious problems especially in divided societies.
(c) It excludes small parties, and hence, easily, minority groups from obtaining a representation.
(d) It spoils many votes, possibly even the majority of them.
(e) As typically there will be two candidates, both close to the center of the political spectrum, many voters will not find a suitable candidate, mostly if the spectrum is large.

In addition, some of the advantages may actually be considered flaws.

(f) A one-party government may exclude the possibility of the establishment of an effective "social contract" among contrasting groups of interest.
(g) An extremist party is possibly more dangerous if it is excluded from the Parliament.
(h) A candidate may be locally strong not due to her/his moral and intellectual standing but to her/his link with a lobby or a powerful pressure group (or person).
(i) There will be a limited competition within the large partitions of the electorate, usually "left" and "right," because each partition will candidate just one person in each district, what can reduce the quality of the candidates.
(j) A perceived excess or responsibility may induce inaction.
(k) Inaction may also be the consequence of the necessity of composing contrasting interests in a party forced to look for the sustain of different social groups (as suggested above in section "Proportional Systems")
(l) A proportional system adopting the D'Hondt rule, or even better the Sainte-Laguë rule, may propitiate independent candidacies not less than majority.

Proportionality, Advantages

(a) It is the "natural" system, because it reproduces the overall society in a manageable assembly (see above, section "Majority, Proportionality and History").
(b) It creates homogenous parties, thus avoiding the problems referred to above under letter *k*.
(c) It puts the political transactions more in the open, as it will be more *among* parties than *within* parties.
(d) It propitiates the representation of minorities.
(e) It offers the voter a broader choice, thus enhancing the competition among parties.
(f) It allows more stability and less traumatic changes, as the governments will typically be made by coalitions.
(g) It strengthens the link between citizens and politicians, as it offers the voters a larger choice.

Proportionality, Flaws

(a) Coalition governments may induce a stalemate in the activity of the government.
(b) A too fragmented system of parties can make the political bargaining cumbersome.
(c) Small centrist parties will be given an excess of power, what reduces the actual proportionality of the system.

Again, some advantages may actually turn into pitfalls.

(d) The stability may correspond to an excessive entrenchment in power of large centrist parties.
(e) Extremist parties may enjoy a undeserved status.

The preceding summary indicates that the typical features of the two meta-systems may easily be either an advantage or a flaw, according to the circumstances, what, arguably, makes it difficult to assess their nature on a general basis. Consequently, the support for either system is usually lexicographic. Those who prefer majority claim that the most relevant advantage of the system, i.e., the enhanced governability given by a reduction of the deciding subjects, is sufficiently important to make it preferable despite its flaws, while those who favor proportionality claim the same with reference to representativeness.

Majority or Proportionality? Evidence

A sound quantitative comparative assessment is made difficult, if not impossible, by the lack of sufficient data. The econometric literature is vast, yet its results are contradictory, as they rest too strongly on the choice of data and on unavoidable simplifying assumptions. However, on a less sophisticated basis data offer only a limited support to the basic argument in favor of majority. Already a couple of decades ago, some authors, like Lijphart (1999) and Farrell (1997), noticed that there is no convincing evidence of the superiority of plurality in term of governability. In Europe, at the beginning of 2017, the two large nonproportional countries (UK and France) apparently faced serious problem of governability, and this is also the case of the three less proportional countries among the proportional ones, that is Spain, Italy, and Greece. Actually, two main tenets in favor of the argued greater governability of majority/plurality systems are not confirmed by facts. The first is the blackmailing power of small centrist parties: in the two more proportional countries of Europe, Italy (1948–1993) and the Netherlands (since 1946), the possibility for a small party to blackmail a majority through the menace of leaving has been a very rare occurrence. The likely reason is that a power rent induces the birth of further small centrist parties, until the rent is dissipated, as first suggested by McGann, Ensch, and Moran in 2009 and as confirmed by data.

As a result, quite a consensus has been reached among scholars that proportionality is preferable, as we noted in section "Majority, Proportionality and History" above, yet this consensus is far from being unanimous, and the debate is far from been conclusive on what proportional system should be deemed the best one. Probably a plurality of scholars would suggest a list proportional system with a threshold. (A recent contribution pointing to this result is Raabe and Linhart 2017; albeit, as we noticed, cross-country comparisons are risky, the result is strengthened by having the authors taken into account a very large set of data, 590 elections in 57 countries.) As a general conclusion, however, it is very likely that no system may be assumed to be the best one for all the existing democracies, because the specific features play a very relevant role. Probably there are systems which are out of date or inferior to others in many if not most cases, yet Katz (1997, p. 308) has surely a point when he claims that what is the best electoral system depends on "who you are, where you are, and where you want to go."

Electoral Reforms and Public Opinion

Broadly speaking, an electoral reform may be either the result of an attempt to make the system closer to the expectations of the citizens or on the contrary of an attempt of the political élite in power to escape a greater control by the citizens, be it to defend some privileges or to enhance the governability. For instance, the first is the case of New Zealand, that in 1994 moved from plurality, as in the UK, to mixed-member proportionality, as in Germany; the second is that of Italy, where two successive electoral laws, in 2014 and in 2016, have been rejected by the Constitutional Court. Dissatisfaction with the political establishment makes it often appealing for the public opinion the proposal of a change in the electoral system. Survey data show that in the UK a majority would prefer to change to proportionality, but in Italy in 1993 a referendum introduced a parallel system, with a larger share for plurality, in lieu of the nearly perfect proportionality in force. (The proportionality was reintroduced in 2005.) Not surprisingly, the electoral system is more likely reformed in times of economic or social change. A source (Renwick 2011) lists 47 modifications of the electoral law across Western Europe from1945 to 2008; only 5 occurred between 1960 and 1980.

It is difficult to disentangle the satisfaction (or the dissatisfaction) with overall politics and politicians from that with the electoral system, not to speak of the low number of observations available; hence the data on the attitude of the citizens are inconclusive. Yet, if we admit that the turnout may be a proxy for the relative approval of the citizens for their electoral system, there are hints that proportionality is more appreciated than plurality. This is what is suggested by the last

elections (at the end of 2016) in comparable countries of Europe. The turnout was 75% in Italy in 2013 and in the Netherlands in 2012, 71.5% in Germany in 2013, 66% in the UK in 2015, and 57% in France in 2012. The first three countries are proportional; the last two are not. That voter turnout is higher in proportional elections is also claimed by IDEA (2016, p. 36), as an overall result of the literature. Among others, Dalton (2008) confirms this guess at the textbook level. In addition, Renwick (quoted above) reports that out of the 84 changes of the electoral law made in Europe (including non-Western countries) between 1945 and 2008, 37 aimed at increasing the proportionality and 23 at reducing it (the remaining ones did not affect this side of the electoral rule).

References

Ahmed A (2013) Democracy and the politics of electoral systems choice. Cambridge University Press, New York

Bowler S, Farrell DM, Petitt RT (2005) Expert opinion on electoral systems: so which electoral system is "best"? J Elections Public Opin Parties 15:3–19

Dalton RJ (2008) Citizen politics. CQ Press, Washington, DC

Diss M, Doghmi A (2016) Multi-winner scoring election methods: Condorcet consistency and paradoxes. Public Choice 169:97–116

Farrell DM (1997) Comparing electoral systems. Macmillan, Basingstoke

IDEA (International Institute for Democracy and Electoral Assistance) (2005) Electoral system design: the new international IDEA handbook. IDEA, Stockholm

IDEA (International Institute for Democracy and Electoral Assistance) (2016) Voter turnout trends around the world. IDEA, Stockholm

Katz RS (1997) Democracy and elections. Oxford University Press, New York

Lijphart A (1999) Patterns of democracy. Government forms and performance in thirty-six countries. Yale University Press, New Haven

McGann A, Ensch J, Moran T (2009) The surprisingly majoritarian nature of proportional democracy. Paper presented in the annual meeting of the APSA, Toronto

Merrill S III (1984) A comparison of efficiency of multicandidate electoral systems. Am J Polit Sci 28:23–48

Migheli M, Ortona G, Ponzano F (2014) Competition among parties and power: an empirical analysis. Ann Oper Res 215:201–214

Raabe J, Linhart E (2017) Which electoral systems succeed at providing proportionality and concentration? Promising designs and risky tools. Eur Polit Sci Rev. Published on line Jan, 17

Renwick A (2011) Electoral reform in Europe since 1945. West Eur Polit 34:456–477

Further Reading

Farrell DM (2011) Electoral systems: a comparative introduction, 2nd edn. Palgrave Macmillan, Houndmills, Basingstoke

Gallagher M (ed) (2005) The politics of electoral systems. Oxford University Press, Oxford, UK

Norris P (2004) Electoral engineering. Cambridge University Press, Cambridge, UK

Emissions Trading

Edwin Woerdman
Faculty of Law, Department of Law and Economics, University of Groningen, Groningen, The Netherlands

Abstract

Emissions trading is a market-based instrument to achieve environmental targets in a cost-effective way by allowing legal entities to buy and sell emission rights. The current international dissemination and intended linking of emissions trading schemes underlines the growing relevance of this instrument. There are three basic design variants of emissions trading: cap-and-trade (allowance trading), performance standard rate trading (credit trading), and project-based credit trading (such as domestic offsets, JI, and the CDM). These design variants are analyzed in terms of effectiveness, efficiency, and acceptance. It is also explained why emissions trading schemes may become inefficient hybrids of such design variants.

Synonyms

Allowance trading; Cap-and-trade; Credit trading; Permit trading; Tradable emission rights; Transferable discharge permits

Definition

Emissions trading is a market-based instrument to achieve environmental targets in a cost-effective way by allowing legal entities to buy and sell emission rights.

Introduction

Emissions trading is a market-based instrument to achieve environmental targets in a cost-effective way by allowing legal entities to buy and sell emission rights. The term "emissions trading" is an umbrella concept for several design variants that differ substantially both in theory and in practice.

In one system, trade in emissions is carried out by firms subject to an absolute emissions cap, which is referred to as "cap-and-trade" or "allowance trading." Examples are the US sulfur dioxide (US SO_2) emissions trading program, the European Union Emissions Trading Scheme (EU ETS), and the California Cap-and-Trade Program to cost-effectively reduce greenhouse gases, such as carbon dioxide (CO_2), as well as some of the Chinese CO_2 emissions trading pilots, including the one in the province of Guangdong.

The other system is based on trade in emissions carried out by firms subject to a relative emissions standard, which is referred to as "performance standard rate trading" or sometimes "credit trading." Think of the early emissions trading schemes in the USA to flexibly maintain air quality standards, the (recently terminated) nitrogen oxide (NO_x) emissions trading scheme in the Netherlands, or the CO_2 emissions trading pilot in the Chinese city of Shenzhen.

Again another system is project-based credit trading, where an investor receives credits for achieved emission reductions at a domestic or foreign host. These emission reductions are measured from a baseline that estimates future emissions at a project location if the project had not taken place. Think of domestic offsets or the international projects of Joint Implementation (JI) or the Clean Development Mechanism (CDM) under the Kyoto Protocol on climate change.

This article is organized as follows. Section "Emissions Trading: Theory and Practice" discusses the concept and international dissemination of emissions trading. Section "Emissions Trading Variants" describes the aforementioned three basic emissions trading variants in more detail. Section "Analysis of Emissions Trading Variants" analyzes these emissions trading variants in terms of effectiveness, efficiency, and acceptance. Section "Emissions Trading Hybrids: the Case of the EU ETS" focuses on hybrids of the emissions trading variants by taking the EU ETS as an example. A summary is presented in section "Summary".

Emissions Trading: Theory and Practice

In the previous century, law and economics scholar Ronald Coase (1960) postulated that trading emissions would improve the cost-effectiveness of environmental regulation. The economic concept of emissions trading was developed further by John Dales (1968). To explain this concept in an easy way, one could say that emissions trading resembles a waterbed. Suppose you would like to raise the water level at the head of a waterbed, then you would have to push down the water level at the foot of the bed. This is also how emissions trading works. One company can buy emission rights and may emit more, but the company selling the emission rights first has to reduce its emissions by an equal amount in order to make these rights available for sale. As a result the government can be certain, provided there is adequate monitoring and enforcement, that the emissions of all the companies together will not exceed the number of emission rights allocated in the emissions trading scheme. As a consequence, emissions trading is an effective tool for achieving emissions targets. Emissions trading is also efficient in the sense that companies can look for the cheapest way to fulfill their emission reductions obligations. The specific design of an emissions trading scheme determines its environmental effectiveness, economic efficiency, and political acceptability.

Emissions trading has gone truly global. Emissions trading schemes have emerged in North America, including the Regional Greenhouse Gas Initiative (RGGI) and the Western Climate Initiative (WCI) in the USA as well as the Québec Cap-and-Trade System in Canada. In Europe, there is the EU Emissions Trading Scheme (EU ETS) for greenhouse gases as well as the Swiss Emissions Trading Scheme. Oceania holds Australia's Carbon Pricing Mechanism (AUS CPM) and the New Zealand Emissions Trading Scheme (NZ ETS). Asia has the Tokyo Cap-and-Trade Program, seven emissions trading pilots carried out in China (for instance, in Beijing and Shanghai), a forthcoming Korea Emissions Trading Scheme, as well as a pilot in Kazakhstan. Moreover, Turkey, Ukraine, and the Russian Federation are considering the adoption of emissions trading schemes, as well as countries such as Thailand and Mexico (for an overview, see, for instance, ICAP 2014).

An interesting, recent development is the intended linking of some of these emissions trading schemes (Weishaar 2014). In North America, the California Cap-and-Trade Program has been linked to the Québec Cap-and-Trade System in 2014. The EU and Australia agreed on a pathway for linking the EU ETS and the AUS CPM in 2018 (although the latter scheme has been under fire in domestic Australian politics). The EU is also negotiating with Switzerland on linking the EU ETS with the Swiss ETS.

Emissions Trading Variants

There are three basic design variants of emissions trading:

- Cap-and-trade
- Performance standard rate trading
- Project-based credit trading

Cap-and-Trade

The cap-and-trade system (or allowance trading) imposes a cap on the annual emissions of a group of companies for a number of years. The emission rights are allocated to established companies for the entire period either for free or through annual sale by auction (a combination is also possible). Newcomers and companies seeking to expand must purchase rights from established companies or from a government reserve, while a company closing down a plant can sell its emission rights.

Some examples are the US sulfur dioxide (US SO_2) emissions trading program initiated in 1995, the European Union Emissions Trading Scheme (EU ETS) for greenhouse gases set up in 2005, and some of the Chinese CO_2 emissions trading pilots, for instance, in the province of Guangdong, erected in 2013.

To introduce and start a cap-and-trade scheme, the legislator has to put in place the following design elements (Tietenberg 1980; Nentjes et al. 2002):

- Set a cap on total emissions per year for a group of emission sources, in advance, and for a range of successive years.
- Create allowances, entitling emissions equal to the total emissions cap.
- Distribute the allowances among group members, either by auctioning the allowances at predetermined dates or by handing out the allowances free of charge.
- Allowances can be traded freely.
- Monitor emissions per source and track the allowances.
- Check compliance over a past budget period (e.g., a book year) by comparing monitored emissions with the number of allowances handed over at the end of the period.
- Penalize noncompliance if emissions exceed the allowances, with a fine per lacking allowance that is a multiple of the market price of an allowance.

Performance Standard Rate Trading

Performance standard rate trading (or credit trading) is different from cap-and-trade. A system of performance standard rate trading is based (not on an emissions cap but) on a mandatory emissions standard adopted for a group of companies. The emissions standard dictates permitted emissions per unit of energy consumption or per unit of added value. In this system emission reduction

credits can be earned by emitting less than what is prescribed by the emissions standard. These credits can then be sold to companies that can use them to compensate their emissions in excess of the emissions standard which applies to them.

If the economy grows, the supply of credits also increases because companies do not operate under an absolute emission ceiling but have to observe a relative emissions standard. An energy-intensive company that expands production, or a newcomer entering the industry, therefore has a right to new emissions, as long as it obeys the emissions standard. This means that absolute emissions will grow. To prevent that from happening, the emissions standard can be strengthened (Dewees 2001; Weishaar 2007).

This system has been developed since the 1970s in the US Environmental Protection Agency (EPA) emissions trading program. A system of tradable nitrogen oxide (NOx) emission reduction credits has been in place for energy-intensive companies in the Netherlands between 2005 and 2013. A system of tradable reduction credits can also be found in the CO_2 emissions trading pilot in the Chinese city of Shenzhen, launched in 2013.

Project-Based Credit Trading

Project-based emissions trading, like Joint Implementation (JI) and Clean Development Mechanism (CDM) projects under the Kyoto Protocol, can be seen as a variant of credit trading. JI relates to emission reduction projects in East European countries, whereas the CDM refers to such projects in developing countries. Both credit trading and emission reduction projects allow for the transfer of credits, but projects usually require pre-approval to check the environmental integrity of the project baseline. This is not necessary under credit trading where the baseline is existing environmental policy (like an emissions standard), so that compliance can be checked by the end of the year. Moreover, the firm which funds the reductions under credit trading is also the firm where the reductions are realized, but in the case of project-based emissions trading, three design options are available (Dutschke and Michaelowa 1999):

- Multilateral approach
- Bilateral approach
- Unilateral approach

In the multilateral approach, an international fund would be created in which private and/or public entities from industrialized countries are required to pool their investments. The bilateral model places more emphasis on private investment and market forces since project selection and implementation are left to the participants. In the unilateral model, a (legal entity within the) host government generates the credits on its own without foreign direct investment.

If such projects are not applied internationally, but domestically, the literature talks about "domestic offsets." A domestic party then invests in a domestic project in order to generate credits that can be used by the former to meet certain emission requirements.

Project-based credit trading generates credits on the basis of the difference between baseline emissions and predicted (or actual) emissions at the project site. The baseline is an estimation of future emissions at the project site in the absence of the project. This baseline is thus a counterfactual that will never materialize.

Analysis of Emissions Trading Variants

What are the similarities and differences between cap-and-trade, performance standard rate trading, and project-based credit trading, and how do they perform in terms of effectiveness, efficiency, and acceptance? This is explained and discussed in the next subsections.

Effectiveness

Effectiveness refers to reaching the emissions targets in an emissions trading scheme. Cap-and-trade is environmentally effective because it imposes an absolute limit on total emissions. If monitoring and enforcement are in order, the emissions cap will be achieved. This is a very strong and important property of cap-and-trade. Effectiveness will only be a problem if emissions are not adequately monitored or if noncompliance

measures are not enforced. Emissions will then be higher than intended by the legislator. However, if the institutional capacity of a country is solid, cap-and-trade will reach effectiveness.

Performance standard rate trading may have a problem in reaching an absolute emission level for the industry to which the relative standard applies. As production and energy consumption rise, the emissions of companies bound by the emissions standard rise proportionally. This will not change if tradable reduction credits are added to the emissions standard. The only way to circumvent this problem is by strengthening the emissions standard. Firms could lobby against such a sharper emission requirement, unless the adjustment to the standard is automatically made.

Project-based credit trading is likely to invoke even stronger effectiveness concerns. There may be several plausible ways to calculate the baseline for emission reduction projects, in particular in host countries where domestic climate change policy is being developed or where such policy is absent. The problem is that the baseline emissions which would have occurred without the project will never be known when the project is implemented. Effectiveness can be undermined if future emissions are overestimated by inflating the baseline to claim more credits. This incentive is strongest for the investor and host under the CDM in developing countries that do not have a nationwide emissions cap. Even if credits are generated on the basis of genuine emission reductions achieved at the project location, emissions may still increase in the CDM host country outside this location (Jepma and Munasinghe 1998). The incentive to inflate the project baseline also exists for legal entities involved in a JI project in Eastern Europe, but not for the JI host party government with an assigned amount of emissions under the Kyoto Protocol since this government would run the risk of being in noncompliance by transferring too many credits.

Efficiency

Efficiency refers to the pricing of emissions so that consumers internalize the environmental damage. Sometimes efficiency is interpreted more modestly as cost-effectiveness, which is about reaching the environmental targets at the lowest possible cost.

Cap-and-trade has "superior" efficiency properties (Tietenberg et al. 1999). In an emissions trading system based on emissions caps, each emission unit has a price and reductions in these units are profitable. From an efficiency point of view, it does not matter whether the emission allowances are allocated for free or sold at auction. If the company uses the free rights to cover its emissions, so-called opportunity costs are associated with them (Nentjes et al. 2002; Woerdman et al. 2008). The company then foregoes the opportunity to sell its allowances and misses sales revenues. The opportunity costs constitute a part of the cost price of the product. In a cap-and-trade system, each unit of emission has a price. Consequently, emission reductions in a cap-and-trade system are profitable regardless of the method by which they are achieved. It makes no difference whether such a reduction is achieved through cleaner exhaust gases, through cuts in energy consumption, or by limiting production. In a cap-and-trade scheme, there is an incentive to examine all emission reduction possibilities and to apply the least-cost option.

In a system of performance standard rate trading, companies that would have to make high costs to achieve the emissions standard will instead buy emission reductions from companies able to comply with the emissions standard at a lower cost. This improves cost-effectiveness, but when one compares such tradable reduction credits to cap-and-trade, then credit trading contains an important inefficiency. Although the emissions standard limits the emissions, the emissions within the limits set by the emissions standard remain without a price (Nentjes and Woerdman 2012). When selling credits, the received amount of money is equal to the sum paid by companies that exceed the emissions standard to purchase the credits. Consequently, for the group of companies as a whole, the cost of the permitted emissions is nil. Pollution not exceeding the relative standard is for free, and absolute emissions are allowed to rise if production or energy consumption increases.

As a consequence, there is no incentive in a performance standard rate trading system to reduce emissions by economizing on fuel input or by slowing production, because this does not earn emission reduction credits. Such credits can only be earned through reducing emissions per unit of energy or output. Mandated emissions released in producing the good have no price, and therefore, their costs are not included in the price of the product. The price of the product is then too low, which leads to overconsumption. The wider range of reduction possibilities in a cap-and-trade system leads to lower total emission reduction costs than in a performance standard rate trading system (Dewees 2001; de Vries et al. 2013).

An alternative, perhaps more simple way of explaining the difference in efficiency properties between cap-and-trade and credit trading is by looking at the supply and demand of emission rights (Woerdman 2004). If the economy grows in a cap-and-trade scheme, the demand for allowances increases, but the supply of allowances remains constant as a result of the emission ceiling. The emissions target will be achieved, and the emissions scarcity is reflected in a higher price for carbon-intensive products. If the economy grows in a credit trading scheme, however, not just demand but also supply of credits will increase since companies do not have an emission ceiling but have to observe an emissions standard. If an energy-intensive company wants to expand production, or if a newcomer enters the industry, it thus has a right to new emissions. The consequence is that the social costs of the extra emissions are not fully reflected in the costs per unit of product. Carbon-intensive products are therefore priced too cheaply.

Project-based credit trading improves the cost-effectiveness of reaching the emissions targets of the investor, but certainly does not have the full-blown efficiency properties of cap-and-trade where each unit of emissions has a price. The project-based variant of trading emissions also suffers from relatively high transaction costs, such as information costs, contract costs, and enforcement costs. Projects such as those under the JI or CDM framework usually require pre-approval to check the environmental integrity of the project baseline, thereby raising transaction costs. Baseline standardization could improve this, by means of developing business-as-usual scenarios for several project types and regions, so that it will not be necessary anymore to construct a baseline for each individual project.

Acceptance

In order to be implemented, an emissions trading scheme needs to be politically acceptable. If an emissions trading scheme is not accepted by a political majority (or by a powerful minority), there will be no emissions trading scheme in the first place.

With respect to acceptance, the emissions trading literature has mainly focused on acceptance by companies (Dijkstra 1999). Performance standard rate trading produces cost savings for both buyers and sellers. The introduction of these credits, to supplement an emissions standard, is therefore likely to receive broad support from companies. The sale by auction of emission rights under an emissions cap is likely to be less acceptable for companies, in particular for those firms that compete on an international product market (Grubb and Neuhoff 2006). Allocating allowances for free basically has the same efficiency properties as selling the allowances at auction (Hahn and Stavins 2011).

Established companies that grow relatively fast are aware that they will not be permitted to produce higher emissions under a cap-and-trade scheme as they expand their production capacity. For this reason, they do not only prefer free rights above sale by auction, but they would rather opt for performance standard rate trading. However, companies that are able to cut emissions more cheaply prefer cap-and-trade because the demand for emissions in this system is higher than in a credit trading scheme.

There is some literature that has also focused on acceptance by politicians (Woerdman 2004). Despite the efficiency advantages of allowance trading, politicians are usually still somewhat inclined to opt for credit trading. A political economy explanation is that credit trading has advantages for certain interest groups, such as the

energy-intensive industries which do not have to purchase extra emission rights if they want to expand their production. However, there are also some advantages of credit trading for politicians themselves. Allowance trading sets emission ceilings via a rather "complex" process of allocating environmental property rights, whereas credit trading more "simply" uses existing environmental policy to calculate the tradable emission reductions. The "political transaction costs" (or start-up costs) of allowance trading are relatively high since it comes to replace existing environmental policy, while credit trading builds increasingly on extant policy, ineffective and inefficient as it may be. Moreover, under allowance trading, a choice must be made between auctioning allowances or giving them away for free, whereas emissions are always handed out for free under credit trading.

The acceptance of project-based emissions trading depends on a number of things, including the political desirability of an emissions cap, the monitoring and enforcement capacities of the administrative infrastructure, and the political merit of using market-based instruments in the first place. Countries that do not want to impose an emissions cap on their industries now, for instance, because they do not want to restrain economic growth in order to fight poverty, will not implement a cap-and-trade scheme. They are also likely to prefer project-based emissions trading to performance standard rate trading, the latter of which is more systematic in approach and thus more demanding in terms of administrative infrastructure. Governments that are critical of neoliberal market approaches are less likely to adopt any of those emissions trading design variants, although their views could change if they discover that they can earn money from selling emission rights. This attitude change has been witnessed for many poor developing countries in the case of the CDM under the Kyoto Protocol, for instance, in Africa. Being opposed first to rich countries "buying their way out" of their reduction obligations, poor African countries actually wanted to attract more CDM projects later, as soon as it became clear to them that such projects are lucrative and that most projects were going to countries like China where transaction costs are lower.

Emissions Trading Hybrids: The Case of the EU ETS

The European Union Emissions Trading Scheme (EU ETS) is a hybrid of the emissions trading design variants discussed above (Nentjes and Woerdman 2012). This regional carbon market has been up and running since 2005 and targets the power sector, a number of industrial sectors, and the aviation sector. It covers 30 countries in Europe and caps about 40% of their greenhouse gas emissions. The EU ETS has a number of implementation problems (e.g., Faure and Peeters 2008; Jaraitė et al. 2013; de Perthuis and Trotignon 2013; van Zeben 2014), for which there is not enough space here to treat them at length (for a recent overview and analysis, see Woerdman 2015). One of the causes of those implementation problems (but certainly not the only one) is that the EU ETS is a hybrid in at least two ways:

- By combining cap-and-trade with elements of performance standard rate trading
- By allowing a limited import of credits generated from CDM projects

The first hybrid element of the EU ETS concerns its new entrant and closure provisions, which resemble those of a performance standard rate trading scheme. In the textbook model of cap-and-trade, new and growing firms have to buy allowances from established companies or from a government reserve. The EU ETS, however, allocates allowances free of charge to newcomers as well as to industries expanding their production capacity (in excess of 10%). This is primarily the result of industry lobbying. Moreover, in the textbook model of cap-and-trade, a company closing down a plant can sell the allowances that remain. In the EU ETS, however, allowances need to be surrendered in case of installation closure or in case of significant decline in production capacity. This was mainly desired by politicians considering it unfair if firms would keep their allowances in such cases. Unfortunately, these credit-like rules lead to the following inefficiencies (e.g., Ellerman 2007; Nentjes and Woerdman 2012).

If companies would keep their allowances in case of closure, it would be more attractive to shut down old, inefficient plants since the allowances could then be sold. Since a company loses its allowances under the EU ETS, however, the closure of dated, climate-unfriendly installations is made less attractive, which is inefficient. Moreover, if (variable) production costs cannot be covered anymore, a company would normally shut down its installations and leave the market. Since a company will lose its allowances under the EU ETS, however, an incentive is provided to companies that make losses to maintain production capacity in order to continue receiving free allowances which can then be sold. Maintaining capacity which is not deployed for production purposes is inefficient. In addition, when newcomers or expanding firms make calculations preceding their capacity investment decision, they do not have to take the market value of their allowances into account because they get them for free. No opportunity costs are attached to this because allowances are surrendered in the event of plant closure or decline in production capacity. Here the allowances allocated for free act as credits when deciding on production capacity, like in a system of performance standard rate trading. Carbon therefore remains unpriced when production capacity is expanded, which is inefficient.

The second hybrid element of the EU ETS concerns its possibility to import relatively cheap credits generated from CDM projects to further enhance cost-effectiveness. CDM credits can be traded for EU allowances. To safeguard the effectiveness of the scheme, however, the import of CDM credits is quantitatively restricted for companies (to 11% of their allocation in the period 2008–2012 or, for newcomers, to 4.5% of their verified emissions during the period 2013–2020). The reason for this quantitative restriction is twofold: the import of such credits increases the overall emissions cap and the environmental integrity of project-based credits may be weaker than that of allowances in a cap-and-trade system. There are also qualitative restrictions as CDM credits are not allowed to be imported from nuclear energy projects, reforestation activities, or projects involving the destruction of industrial gases.

Summary

Emissions trading is a market-based instrument to achieve environmental targets in a cost-effective way by allowing legal entities to buy and sell emission rights. There are three basic design variants of emissions trading: cap-and-trade (allowance trading), performance standard rate trading (credit trading), and project-based credit trading (such as domestic offsets, JI, and the CDM). In practice, emissions trading schemes can be hybrids of these design variants.

Cap-and-trade is effective in reducing absolute emission levels because it operates under an absolute emissions cap. Cap-and-trade is also efficient, since each unit of emissions has a price. When emissions start to rise in a performance standard rate trading scheme, however, an absolute emission level can only be achieved if the emissions standard is strengthened. An inefficiency of performance standard rate trading is that not every unit of emissions has a price. Moreover, as opposed to cap-and-trade, there is no incentive to reduce emissions by economizing on fuel input or by slowing production. Project-based credit trading is less systematic and more ad hoc due to its project focus. This also entails relatively high transaction costs, although standardizing baselines for several project types and regions helps to reduce those costs.

Firms will generally lobby in favor of free emission rights based on performance standard rate trading. Various politicians also tend to favor such tradable credits if this builds upon already existing direct regulation. As a result, emissions trading schemes sometimes become inefficient hybrids of the design variants discussed above, such as the EU ETS.

Nevertheless, those politicians that want to reach an absolute emission level at the lowest possible cost favor cap-and-trade. The current international dissemination and intended linking of cap-and-trade schemes underlines the growing relevance of emissions trading. Future research on the design elements and implementation problems of real-life emissions trading schemes could help to improve their effectiveness, efficiency, and acceptance.

Cross-References

▶ Transferable Discharge Permits

References

Coase RH (1960) The problem of social cost. J Law Econ 3:1–44
Dales JH (1968) Pollution, property and prices: an essay in policy-making and economics. Toronto University Press, Toronto
de Perthuis C, Trotignon R (2013) Governance of CO_2 markets: lessons from the EU ETS. Working paper series no 2013–07, CDC Climat, Climate Economics Chair, Paris
de Vries FP, Dijkstra BR, McGinty M (2013) On emissions trading and market structure: cap-and-trade versus intensity standards. Environ Resour Econ. https://doi.org/10.1007/s10640-013-9715-2
Dewees DN (2001) Emissions trading: ERCs or allowances? Land Econ 77(4):513–526
Dijkstra BR (1999) The political economy of environmental policy: a public choice approach to market instruments. Edward Elgar, Cheltenham
Dutschke M, Michaelowa A (1999) Creation and sharing of credits through the Clean Development Mechanism under the Kyoto Protocol. In: Jepma CJ, van der Gaast WP (eds) On the compatibility of flexible instruments. Kluwer, Dordrecht, pp 47–64
Ellerman D (2007) New entrant and closure provisions: how do they distort? Energy J 28:63–78
Faure M, Peeters M (eds) (2008) Climate change and European emissions trading: lessons for theory and practice. Edward Elgar, Cheltenham
Grubb M, Neuhoff K (2006) Allocation and competitiveness in the EU emissions trading scheme: policy overview. Clim Policy 6(1):7–30
Hahn RW, Stavins RN (2011) The effect of allowance allocations on cap-and-trade system performance. J Law Econ 54:267–294
ICAP (2014) International overview of emissions trading schemes. Available online at https://icapcarbonaction.com/ets-map
Jaraitė J, Jong T, Kažukauskas A, Zaklan A, Zeitlberger A (2013) Matching EU ETS accounts to historical parent companies: a technical note. European University Institute, Florence. Available online at http://fsr.eui.eu/CPRU/EUTLTransactionData.aspx
Jepma CJ, Munasinghe M (1998) Climate change policy: facts, issues and analyses. Cambridge University Press, Cambridge
Nentjes A, Woerdman E (2012) Tradable permits versus tradable credits: a survey and analysis. Int Rev Environ Resour Econ 6(1):1–78
Nentjes A, Boom JT, Dijkstra BR, Koster M, Woerdman E, Zhang ZX (2002) National and international emissions trading for greenhouse gases. Dutch National Research Programme on Global Air Pollution and Climate Change (NRP) Report No. 410 200 093, Bilthoven
Tietenberg T (1980) Transferable discharge permits and the control of stationary source air pollution: a survey and synthesis. Land Econ 56(4):391–416
Tietenberg T, Grubb M, Michaelowa A, Swift B, Zhang ZX (1999) International rules for greenhouse gas emissions trading: defining the principles, modalities, rules and guidelines for verification, reporting and accountability. UNCTAD/GDS/GFSB/Misc.6, United Nations Conference on Trade and Development (UNCTAD), Geneva
van Zeben J (2014) The allocation of regulatory competence in the EU emissions trading scheme. Cambridge University Press, Cambridge
Weishaar SE (2007) CO_2 emission allowance allocation mechanisms, allocative efficiency and the environment: a static and dynamic perspective. Eur J Law Econ 24:29–70
Weishaar SE (2014) Emissions trading design: a critical overview. Edward Elgar, Cheltenham
Woerdman E (2004) The institutional economics of market-based climate policy. Elsevier, Amsterdam
Woerdman E (2015) The EU greenhouse gas emissions trading scheme. In: Holwerda M, Roggenkamp MM, Woerdman E (eds) Essential EU climate law. Edward Elgar, Cheltenham (forthcoming)
Woerdman E, Arcuri A, Clò S (2008) Emissions trading and the polluter-pays principle: do polluters pay under grandfathering? Rev Law Econ 4(2):565–590

Empirical Analysis

Alexander J. Wulf
SRH Hochschule Berlin, Berlin, Germany

Synonyms

Empirical law and economics; Empirical legal research; Empirical legal science; Empirical legal studies

Definition

Empirical analysis uses empirical research methods from economics and the social sciences in an attempt to provide answers to research questions in the field of law, i.e., in order to investigate the operative and functional aspects of law and legal consequences. The goal of empirical legal

research is to make a contribution to all subjects and phenomena that are of interest to legal scholars and for which no methods have previously been available. The use of empirical methods in legal science can therefore lead to results that cannot be achieved by traditional law research with the methods at its disposal. The ultimate aim of the approach is to contribute to a systematic understanding of our legal system based on empirical data.

In empirical analysis research methods from the social sciences are used to examine research questions in the legal sciences in order to study the operative and functional aspects of the law and their effects (Baldwin and Davis 2003, p. 880). The increasing trend toward testing hypotheses and questions has been described as "the next big thing" in legal science (George 2005, p. 142), even as a revolution (Ho and Kramer 2013) whose significance has as yet been underestimated (Gordon 1993, p. 2085).

In the USA the research approaches of the social sciences first began to take on greater significance in connection with legal realism. Legal realism grew up in the 1930s and 1940s, mainly in the USA, and was characterized especially by the use of interdisciplinary methods borrowed from the social sciences (Eisenberg 2011, p. 1720). In legal realism it was, for example, assumed that judges' decisions were determined not only by laws, precedents, and general legal principles but also by the judges' social backgrounds and political convictions (Baldwin and Davis 2003, p. 882). For example, social scientific methods were needed to be able to analyze judges' verdicts from this viewpoint. Since these methods are often empirical, legal realism introduced the empirical paradigm into the legal sciences for the first time. However, these early empirical research efforts remained few and far between, and thus legal realism fell short of its empirical potential. Because legal realism brought the empirical approaches into the legal sciences and thus prepared a basis for their acceptance, it is seen as the trailblazer of empirical research in the legal sciences (George 2005, pp. 144–145).

Since the beginning of this century, the main country in which empirical legal research has been practiced to any great extent is the USA (George 2005, pp. 141–142). Today it has a wide forum (Eisenberg 2011, pp. 1713–1714) in a number of journals with international repute (e.g., the *Journal of Empirical Legal Studies*, the *Journal of Legal Studies*, the *Journal of Law and Economics*, and the *European Journal of Law and Economics*), at conferences (e.g., the Annual Conference on Empirical Legal Studies and the Annual Conference of the European Association of Law and Economics), academic societies (e.g., the Society for Empirical Legal Studies, the American Law and Economics Association, the European Association of Law and Economics), and university research centers (e.g., the University of California at Berkeley – Center for the Study of Law and Society; University of California, Los Angeles – School of Law Empirical Research Group; and Hamburg University – Institute of Law and Economics, Germany). It appears likely that its initial rapid growth will continue in the future. The success of empirical methods in the legal sciences is partially due to the fact that it brings together scholars from several different disciplines who have been working independently of each other on different aspects of the legal system, e.g., the sociology of law ("Law and Society") and the economic analysis of law ("Law and Economics"). This promotes interdisciplinary legal research (Eisenberg 2011, pp. 1719–1720 and 1722–1724).

The common goal of these different disciplines is to develop a systematic understanding of the legal system based on empirical data (Eisenberg 2011, p. 1720). The issues covered by this research are broad. The main focus is on analyzing how legislators enact and implement legal regulations and how these regulations affect the behavior of those to whom they are applied (Ulen 2008, p. 74). Empirical legal research strives as far as possible to make a contribution to all issues and phenomena which, while they are of importance for lawyers, have not previously been empirically investigable because of the lack of methods (Suchman 2006, p. 2).

The development of empirical legal research as an independent field of legal science is sometimes seen as a response to the fact that traditional

scholars of legal doctrine neglected to do empirical research on the actual functioning of the legal system (Eisenberg 2011, pp. 1734–1735). On this view, as a result of the lack of empirical data on our legal systems, many of the theories, assumptions, and prognoses on which traditional legal research has been based, either implicitly or explicitly (Ho and Kramer 2013, p. 1202), have been left without empirical support and therefore remain vague. Moreover, as a result of this neglect, the actors involved in the legal system such as the courts, parties to actions, lawyers, political decision-makers, and society in general are only insufficiently informed about the functional aspects of the legal system (Eisenberg 2011, pp. 1736–1737). Empirical legal research can be seen as an attempt to remedy this state of affairs.

The Contribution of Empirical Analysis to Legal Science

The use of empirical methods in the legal sciences can lead to insights that traditional legal research is unable to achieve with its methods. In this section the limitations of empirical research will be discussed. For reasons of space only a few central aspects will be highlighted.

Objectivity and Value Freedom

Economists and empiricists working in the legal sciences frequently adopt the view, which originated in the natural sciences, that empirical research is objective and can thus be value-free (Friedman 1953, p. 3; see also Sen 1981). However, it is highly doubtful whether it is really possible in the context of economics and the social sciences to explain, analyze, and interpret data in a value-free way. If we assume that it is not possible, this empiricist view reduces the value of empirical research for institutions, decision-makers, and politics (Hausman and McPherson 2006). This may also apply to the legal sciences since this view may result in a neglect of moral and ethical issues. Empirical research data would then be of little use to legal scholars (Hausman and McPherson 2006, pp. 291–308). Thus the support provided by empirical research for finding solutions to normative issues, which requires a weighing up of philosophical and ethical arguments, is limited (Lawless et al. 2010, p. 21).

Qualitative and Quantitative Research Approaches

Since empirical legal science is mainly US based, it uses almost exclusively quantitative research methods, i.e., methods which are used to analyze statistical data (Suchman 2006, p. 2; Chambliss 2008, pp. 25–26). To date qualitative research methods, which are widespread in the social sciences in Europe and are employed to evaluate data from sources such as texts and audio and video and other forms of visual material, have not played such an important role (Baldwin and Davis 2003, pp. 891–892). One reason for this neglect is that in the USA empirical legal science is closely allied to the economic analysis of law ("Law and Economics"). This movement originated in economics which, unlike other social sciences, does not use qualitative research methods on principle, since they are seen as being too "soft," i.e., not sufficiently objective. Since in Europe the sociology of law has traditionally played an important role, its quantitative orientation may have constituted an additional barrier to the spread of empirical analysis in Europe (Posner 1997, pp. 5–6).

Methodological Problems and Oversimplification of the Complexity of Legal Issues

Like all interdisciplinary research approaches, research in empirical legal science requires a high level of specialized knowledge of the various scientific disciplines. The fact that legal scholars need to acquire the methodological know-how of the social sciences (Eisenberg 2011, pp. 1728–1729) and social scientists, the detailed knowledge that legal scholars have of the functioning of the legal system (Baldwin and Davis 2003, p. 883) results in two main problems for empirical analysis. These are the methodological problems of empirical research and the problem of reducing legal complexity. Both problem areas are briefly described below.

Methodologically, empirical legal science is criticized for being of lower-than-average quality, although it can be doubted whether the standards of other forms of (interdisciplinary) research are actually any higher. The reasons for such methodological problems vary. They frequently arise from the incorrect use of methods adopted from the social sciences or from flawed research designs. In order to counteract these problems, empirical legal science needs to develop its own methodological discourse. In the long term this could promote a sustainable awareness of methods and thus improve the quality of empirical analysis (Eisenberg 2011, pp. 1730–1731). Some of the traditional publications of legal science, e.g., in the USA the student-edited law journals and in Europe-edited volumes, may not really be suitable for ensuring the methodological quality of the empirical studies submitted for publication (Chambliss 2008, pp. 26–28).

Another criticism frequently leveled at the empirical legal approach is that they reduce the complexity of legal issues (Baldwin and Davis 2003, p. 883). Such reduction is foreign to classical legal scholars, since it is not required for their traditional methods. However, if the behavior of relevance to law is to be investigated with sufficient methodological rigor in empirical studies, it is often unavoidable. Thus the decision as to what extent basic legal conditions should be taken into account in an empirical study is of substantial importance. If the legal facts are oversimplified in the interests of methodological stringency, the results of such research can have little relevance for legal science. In order to avoid this, it is necessary first to present the legal issue to be examined in its full complexity. Only then can the complexity be reduced as required. Care must be taken to fully inform the recipient of the research results of this reduction (Faust 2006, pp. 849–850). If the complexity is reduced in accordance with these requirements, this usually satisfies the demands of legal science, particularly since it is rare for the results of empirical research to replace legal discourse; as a rule they merely provide an additional perspective.

Finally, it is important to be aware that there are areas of legal science in which such reductions in complexity are not possible and which cannot therefore be investigated by means of quantifying methods. Even technically well-implemented empirical projects must fail when their subject matter is of an ethical and moral nature. Here it can be perturbing if issues that cannot be quantified are more or less uncritically functionalized. It is the task of discourse in legal science to critically assess the value of such research.

Future Directions

In continental Europe empirical legal research is currently less widespread than in the English-speaking countries, and many obstacles to its gaining more acceptance still need to be overcome. Legal scholars who have no basic knowledge of the empirical research methods required may initially find it disturbing that the utility and scope of such research efforts in the legal sciences have often not been sufficiently clearly defined (Hull 1989, p. 915). Moreover, since it is often seen as being closely linked to economics, legal scholars sometimes think that empirical legal science is theoretical and abstract, naively functionalist, or politically conservative. Today these preconceptions can no longer be considered justified (Klerman 2002, p. 1167). Empirical analysis can contribute to a systematic, empirically based understanding of our legal system. It can therefore be assumed that it will play an increasing role in the future, not only in the English-speaking countries but also in continental Europe and other non-English-speaking countries.

Cross-References

▶ Causation
▶ Impact Assessment

References

Baldwin J, Davis G (2003) Empirical research in law. In: Cane P, Tushnet MV (eds) The Oxford handbook of legal studies. Oxford University Press, Oxford, pp 880–900

Chambliss E (2008) When do facts persuade. Some thoughts on the market for empirical legal studies. Law Contemp Probl 71(2):17–39

Eisenberg T (2011) The origins, nature, and promise of empirical legal studies and a response to concerns. Univ Ill Law Rev 2011(5):1713–1738
Faust F (2006) Comparative law and economic analysis of law. In: Reimann M, Zimmermann R (eds) The Oxford handbook of comparative law. Oxford University Press, Oxford, pp 837–865
Friedman M (1953) The methodology of positive economics. In: Friedman M (ed) Essays in positive economics. University of Chicago Press, Chicago, pp 3–43
George T (2005) An empirical study of empirical legal scholarship. The top law schools. Indiana Law J 81(1):141–161
Gordon RW (1993) Lawyers, scholars, and the "middle ground". Mich Law Rev 91(8):2075–2112
Hausman DM, McPherson MS (2006) How could ethics matter to economics. In: Hausman DM, McPherson MS (eds) Economic analysis, moral philosophy, and public policy, 2nd edn. Cambridge University Press, Cambridge, pp 291–307
Ho DE, Kramer L (2013) The empirical revolution in law. Stanf Law Rev 65(6):1195–1371
Hull NEH (1989) The perils of empirical legal research. Law Soc Rev 23(5):915–920
Klerman D (2002) Statistical and economic approaches to legal history. Univ Ill Law Rev 2002:1167–1176
Lawless RM, Robbennolt JK, Ulen T (2010) Empirical methods in law. Aspen Publishers, Alphen aan den Rijn
Posner RA (1997) The future of the law and economics movement in Europe. Int Rev Law Econ 17(1):3–14
Sen A (1981) Accounts, actions and values. Objectivity of social science. In: Lloyd C (ed) Social theory and political practice. Clarendon, Oxford, pp 87–107
Suchman M (2006) Empirical legal studies. Sociology of law, or something ELS entirely? AMICI. Newsl Sociol Law Sect Am Sociol Assoc 13(1):1–4
Ulen T (2008) The appeal of legal empiricism. In: Eger T et al (eds) Internationalisierung des Rechts und seine ökonomische Analyse. Internationalization of the law and its economic analysis, Festschrift für Hans-Bernd Schäfer zum 65. Geburtstag, Gabler, Wiesbaden, pp 71–88

Empirical Analysis of Judicial Decisions

Luciana Yeung
Insper Institute of Education and Research, São Paulo, Brazil

Synonyms

Economic analysis of judicial decisions; Judicial behavior; Judicial decision-making

Definition

Evidence shows that, contrary to what believed traditional legal theorists, judges when making decisions are not merely making a pure exercise of interpretation of the law. They are influenced by factors such as panel composition, material and nonmaterial benefits, pressure groups, etc., and may follow distinct trends. This chapter presents some models that explain judicial behavior or judicial decisions. Recent (and not so recent) empirical literature has been providing a rich debate in this still recent discussion.

Introduction: Why Care? Judicial Outcomes and Economic Performance

> If judges have a substantial amount of discretion in deciding cases, then it is important to know the motives and the value systems which influence their exercise of discretion. (C. Hermann Pritchett 1968)

Coase (1960) taught us that courts and court outcomes – i.e., judicial decisions – impact the economy. Several other authors have also shown empirically that while well-functioning courts provide proper environment for productive activities, guarantee contract enforcement, and reduce uncertainties in the economy, malfunctioning courts may deter growth, investments, job creation, and increase insecurity (e.g., Weder 1995; Sherwood 2004). Since judicial decisions are courts' main "product," evaluating "how judges judge" and "what explains judges' decision making" becomes a crucial task.

Although this is a complex job, which has recently been aided by studies in fields such as cognitive sciences and empirical psychology, scholars of different backgrounds (political sciences, law, economics, sociology, etc.) have been debating about this topic for several decades. Pritchett (1968) accounts Charles G. Haines as being one of the pioneers in the study of judicial behavior, with the publication of "General Observations on the Effects of Personal, Political, and Economic Influences in the Decisions of Judges" in a 1922 edition of the Illinois Law Review. Pritchett himself is considered by many, one of the original creators' of this field (Epstein 2016).

From the View of Judicial Decisions as Pure Interpretations of Law to Richard Posner's Nine Theories of Judicial Behavior

One frequent concern among judicial behavior scholars is whether judges are influenced by their preconceptions or ideologies. A long debate over this issue – and not yet finished – puts *legalists* in one side and *realists* in the other. Adepts of legalism would argue that when judges judge, they are purely interpreting the law, in the best manner they can; therefore, bringing law into life is the main judicial job. Realists, on the other hand, do not believe there is a unique and certain manner to interpret the law. Each judge, when deciding, is unavoidably influenced by prior beliefs, prior personal and/or professional experiences – even if they try to strictly follow legal rules. This all may be called one's *ideology*. Realist scholars then are mainly interested in creating good and accurate measures of judges' ideology.

Adept of the realist view of judicial behavior, Posner (2008) would explain judges' personal attitudes in terms of "Bayesian preferences," as defined by Bayes statistical theorem: it shows that future probabilities of a certain kind of behavior – or decision – may be explained by former probabilities of that same behavior. Thus, in order to predict the occurrence of a certain type of behavior/decision by a judge, researchers should evaluate how was his/her decision in the past.

Besides that, Posner summarized and further developed years of theorization in judicial behavior. The author categorized nine theories of judicial behavior: Attitudinal, Strategic, Sociological, Psychological, Economic, Organizational, Pragmatic, Phenomenological, and Legalist.

Attitudinal: The attitudinal theory explains that judges' decisions are mainly reflections of their political preferences or what it is called political ideology.

Strategic: The strategic theory argues that judges' decisions reflect their preoccupations with external factors, such as opinions and pressures coming from peers, other political powers, and the rest of society (public opinion, media, interest groups, etc.).

Sociological: This theory is focused on smaller judging groups and explains why factors such as panel and voting compositions do affect judicial decision.

Psychological: The psychological theory focuses on explaining how one's preconceptions influence decision-making in circumstances under uncertainty. As Posner poses, legal systems, especially (but not only) the one in the USA, is fundamentally characterized by uncertain events, facts, and information.

Economic: The economic theory presents judges as rational, utility maximizers, who constantly behave in response to incentives and constraints. In this case, utility maximizing may be related to desire for leisure, promotion, good reputation, or even, internal feeling of satisfaction. Thus, the economic model may also encompass the strategic and the sociological theories of judicial behavior.

Organizational: The so-called agent-principal problem is the basis of the organizational theory. Mainly, this model considers judges as agents of a principal (the government) and seeks to explain judicial decisions under this perspective.

Pragmatic: The pragmatic model explains that judges are concerned with and do consider the consequences of their decisions. It is also known as the *consequentialism* approach of judicial decision-making.

Phenomenological: Posner explains that "phenomenology studies first-person consciousness – experience as it presents itself to the conscious mind" (p. 40), so this theory relates to the self-consciousness of judges when judging.

Legalist: As briefly explained above, legalism views judicial decision-making as a "pure" interpretation of the law. Legalists believe that judges when deciding at courts are solely impacted by their effort to apply the law, not being disturbed by other influences, especially, any personal preferences or preconceptions of any kind.

Due to the complexity of the phenomenon of judicial decision-making, Posner recognizes that there is no single theory able to explain it entirely. Therefore, the theories above are, actually, complementary and not substitutes.

This entry adopts some of these perspectives (e.g., attitudinal, economic) and less others (e.g., phenomenological, legalist). In subsequent sections, we provide references for empirical evidence corroborating some of these theories, especially the Attitudinal, the Strategic, and the Sociological.

Common Law Versus Civil Law: Unified or Different Models for Judicial Decision?

Early literature on judicial decisions was based on common law systems, as it happened with most of the literature on economic analysis of law. Yet, the analysis of judicial decisions has no absolute boundaries across different legal systems. The reason is simple. Judges judge lawsuits, no matter where, or under which system. In modern days, even common law judges follow constitutions and statutes, and civil law judges, not rarely, adhere to precedents (some instances in mandatory, or quasi-mandatory manner – as lower court judges following higher courts – and some other instances voluntarily). As Schneider (2005) poses, it is unreasonable to believe, as orthodox legal theory assumes, that civil-law courts only apply the law coming from the legislature. Because of that, "propositions asserting substantial differences between common-law and civil-law systems may be overstated" (p. 139).

Thus, one can infer that judicial decision-making may be analyzed as one single phenomenon, independently of the locus where it happens. This does not mean that one might not eventually find out that the outcomes of judicial decisions differ from one country to another or from one system to the other. It may also be possible to find out that judges do have different tendencies in their judicial making depending on where they are. For instances, one might conclude that judges in some countries tend to be more sensitive to "social issues," such as inequality and wealth distribution and that their decisions reflect these concerns. Yet, these results are related to the *different variables affecting the process of judicial decision*; it does not mean that there are different models of judicial decision, one for each legal system. It is also not true that judges, when faced to similar incentives or constraints, behave differently according to their legal origin. As the literature shows, different modes of judicial behavior are caused by different constraints and incentives which judges are faced to. The different models used to explain judicial behavior, as proposed by Posner's "Nine Theories of Judicial Behavior," described above, illustrate this idea.

Some Early Literature on Judicial Behavior and Decision-Making

Focusing on the US Supreme Court, and beginning in 1940s, Pritchett developed several empirical methodologies for the analysis of judicial behavior, among them, "bloc analysis" and "attitude analysis." While later the author himself considered the first one "a rather primitive device" (1968, p. 499), it was a very useful tool for the evaluation of Justices' voting patterns in non-unanimous decisions. Based on observations for more than 20 years, Pritchett found evidence of persistent divergences between Justices based on ideological differences. Along with "bloc analysis," the author employed "attitude analysis." It was one of the first attempts ever in the literature to include personal characteristics – or "personal policy attitudes" – as determinants of stronger than average voting patterns by the Justices. Pritchett' main hypothesis was that judges are influenced by their personal ideologies, and their decisions at courts are not mere interpretations of the "letters of the law." This methodology, which employed a "box score" to register Justices' attitude, became a classical reference in the literature of empirical analysis of judicial decisions.

Tate (1983) created a compendium of different methodologies applied in different studies. His analysis encompassed qualitative and quantitative approaches and was complemented with critical comments. At that time, the author showed optimism in the potentials of "standard statistical methods such as regression analysis" for the study of judicial behavior (p. 74).

Fon and Parisi (2006) developed a dynamic theoretical model to explain judicial decision making in civil law systems. Their focus was not on judges as individuals, but on the result of their action, i.e., decisions and precedents. The authors evaluated the circumstances under which legal rules may be consolidated or may be corroded at courts. In their model, external shocks usually push the system into a dynamic trend, pulling it away *or* towards the point of consolidation of rules. In turn, institutional threshold of *jurisprudence constante* played a crucial role in the definition of this dynamic path.

Factors Impacting Judicial Decision: Empirical Evidence

Several factors may affect judicial decision, besides the manner by which judges interpret the law (as legalists would argue). Those would include internal factors (such as one's ideology) and external factors (such as pressures from public opinion); some may change throughout one's career (again, ideology would be an example, though here we are speaking of ideology in a broad sense), and others are constant for a certain individual (such as gender or race). The outcome – i.e., how judges effectively judge – is a combination of all those factors, and no single one explains it all, all the time. For this reason, empirically measuring the impact of a certain factor is not an easy task. Luckily, much has been advanced in the last decades. Let us review some literature on these.

Ideology

"How much ideology affect judicial decisions" is one of the most common themes in empirical analysis of judicial behavior. Since the 1st half of the twentieth century, Pritchett succeeded in finding evidence that US Supreme Court Justices were influenced by political ideologies when judging (as before). This is what Posner called "the attitudinal theory," which "deploys. . . highly developed apparatus to determine the extent to which actual judicial decisions reflect the attitudes or 'ideology' of the judge rather than the dictates of the 'law'" (Cameron and Kornhauser 2017, p. 536).

As proponents of this theory, Epstein, Landes, and Posner (ELP 2013) point to the fact that there are *ex ante* and *ex post* measures of judges' political ideology. For Supreme Court Justices, the most common *ex ante* measure is the party of the nominating President. But since nominations must be approved by the Senate (and this is true for many countries), some scholars have also used measures of senatorial influence to capture Justices' ideology. Others have used editorials of influential newspaper to capture US Supreme Court nominees' *ex ante* ideological profile. *Ex post* measures, as one would expect, are based on evaluations of Justices' and judges' votes, speeches, and articles. Researchers usually combine qualitative and quantitative analysis of written and oral material produced to infer one's ideological inclinations. ELP (2013) show that "judicial self-restraint [on personal ideology] has long been in decline [since the 1960s]" for US Supreme Court Justices. In other words, the impact of ideology has been growing over time. For other courts in the USA, the authors indicate that ideology also plays a role, although in weaker magnitudes. This means that ideology has proportionally more influence at the top of the judicial hierarchy (Cameron and Kornhauser 2017).

Supreme Courts outside the USA seem also to be influenced by ideology in their judicial decisions. Yeung and Azevedo (2015) use a set of approximately 1700 decisions of the STJ (Superior Tribunal de Justiça), one of the two supreme courts in Brazil (being STJ decisions eventually appealable to STF, the Federal Supreme Court). Their original objective was to evaluate whether this court tended to favor small debtors in contractual disputes involving financial institutions. There is widespread anecdotal evidence that Brazilian courts favor weaker parties and disfavor banks. As an overall result, the authors did not confirm this popular belief. Yet, when analyzing specific variables, evidence of ideological decision emerges. For instances, Justices at the STJ decide differently depending on who stands as the defendant party: if an individual appear as defendant, the debtor tend to be favored

in a significant manner, as compared to the instances in which an enterprise appear as defendant. Justices seems to believe that, faced with financial institutions, individuals need more protection by the law than enterprises do.

Another evidence of ideology was assessed indirectly in this study. Being an appellate supreme court, STJ receives cases from 2nd degree state courts from all over the country. The authors measured whether there was any "regional factor" affecting STJ decisions. The only state with a significant result was the Brazilian southernmost state, Rio Grande do Sul: cases with that origin were consistently reformed by the STJ Justices, and in the direction of *disfavoring* debtors. It was clear that they used their power to "correct" some pro debtor trend by the Southern judges. Curiously, there is a long lasting theoretical discussion behind this fact: historically, the state of Rio Grande do Sul has been known as the birthplace of a judicial movement called *Associação dos Juízes para a Democracia* ("Association of Judges for Democracy"), whose main goal is to promote "social justice," or more precisely, wealth redistribution, by means of the Judiciary. Judges from that state are known to be adepts to this movement and, for this reason, to be more sympathetic to "social issues," and less favorable to the "big capital," as banks and big enterprises. Ideological influences on their decisions are not only undisguised, but in fact, a clear statement. In this sense, Yeung and Azevedo's results find evidence of two ideological impacts: one by judges in the state of Rio Grande do Sul, which is *pro debtor*, and the other, by the Justices at STJ, overall regarding the southern judges as "biased." Both results were consistent and statistically significant.

Further effects have been observed in Justices' ideologic profiles. ELP (2013) distinguish two phenomena: the *ideological drift,* and the *ideological divergence.* The former captures changes in a Justice's *ex ante* ideology since his/her appointment to the Supreme Court. The second refers to changes in the Justice's ideology compared to the ideology of the nominating President. This effect explains why Justices may, sometimes, vote different to what is expected. Ideological divergence may happen gradually (i.e., Justices start at the Supreme Court much aligned with their nominating President, but then divergence increase due throughout the time), or it may exist since the beginning of a nominee's career at the Supreme Court. In this case, it may be caused by the President's lack of information about the nominee's profile, or it may happen because the President had other political compromises when nominating the Justice, which were unrelated to ideological positions.

Abundant are theories and empirical evidence of ideology impacts on judicial decision-making; it is not the only factor explaining judges' behavior. It would be naïve to believe that judges, even Supreme Court Justice, would be able to use their discretionary power to entirely (and only to) pursue their personal ideologies. Several other constraints curtail these motives.

Gender

The impact of a judge's gender on judicial decision has also been studied in several studies; as before, we bring a short selection.

Peresie (2005) finds judges' gender as an impacting factor on decisions at USA appellate courts on disputes over sexual harassment and sex discrimination. Gender acts as a direct impact factor – i.e., female judges favor more frequently the victims of discrimination – and as an indirect factor, through peer effect on panels – i.e., female judges do influence their male counterparts when judging such cases. Peresie finds that panels with female judges tend to favor alleged victims twice as often as panels with only male judges. In this study, gender was more impacting than ideology on judicial decisions.

Farhang and Wawro (2004) find a strong panel effect by women, i.e., female judges tend to influence their male colleagues in the panels. Yet, they find that a second woman on the panel does not have the same effect as the first one. These authors also try to find evidence of race impact, but – differently to the gender factor – they find none, although they are cautious to interpret this latest result.

Boyd et al. (2010) employ propensity score matching and also find significant gender impact

in sex discrimination disputes. Here, as in Peresie (2005), impacts occur both directly (on individual female judges) and indirectly (on male counterparts in panels). Although the authors analyzed 13 types of judicial disputes, only those on sex discrimination were significantly impacted by the judges' gender. Within their list of analysis, there were other types of disputes that are usually gender sensitive, such as abortion and sexual harassment, but on no other issue had gender significant impact.

Other studies bring similar evidence for courts outside the USA. King and Greening (2007) analyzed decisions by the International Criminal Tribunal on cases of sexual assault in former Yugoslavia. They found that female judges tend to punish more severely defendants who assaulted women – in what constitutes a "gender solidarity" between judges and victims. This solidarity seems also to be present when all-male panels analyzed cases involving male victims: sentences for these cases were more than 100 months lengthier than those in which there was at least one female judge on the panel. In France, under the context of child support orders, Bourreau-Dubois et al. (2014) find that when the parents' average offer is reasonable, female judges tend to be more generous than male judges.

Grezzana and Ponczek (2012) analyzed more than 90,000 labor disputes at the Brazilian Superior Labor Court. Overall, the authors find no evidence of gender impact. However, once they control for the object of dispute, impact is evident for cases such as "wage equalization" and "employment and union link." Under these circumstances, female judges tend to favor female litigants (workers), whereas male judges tend to favor male ones. Again, there seem to be some kind of "gender solidarity" between judges and litigants in the Brazilian Superior Labor Court.

Why would judges' gender have impact on the manner by which they decide? Based on previous literature, Boyd et al. (2010) draw 4 accounts affecting judging, either individually or in groups (panels). First, *different voice* is the manifestation of the differences that male and female individuals see and analyze the world and society; basically, this is the individual perspective of things, particularly here, of the cases disputed at courts. *Representational* account is the manifestation of female judges seeing themselves as representatives of all female individuals in society, and specifically, of female litigants in disputes. Female judges would decide in favor of women in cases where there are particular interests for the whole class of women in society. Third, *informational* account posits female judges as having more information that would be valuable for resolving the dispute. Under such circumstances, their male counterparts would benefit from this privileged information, and the effect will be channeled through the panel voting. Last, *organizational* account will actually oversee the gender impact on judicial decisions; the view here is that professional training and institutional rules in the Judiciary are clear and similar enough to minimize any significant differences between male and female judges. All these accounts have been explored, tested, and analyzed by a rich literature on this topic Boyd et al. (2010) provide detailed references on each of these approaches.

Besides gender, there are other factors affecting judicial decisions, which are related to minority groups such as race, ethnicity, religious group, and social background, among others (e.g., an interesting piece of work by Schwartz and Murchison 2016, on the impacts of ethnicity-nationality at the Constitutional Court of Bosnia-Herzegovina). Due to the limitations of this chapter, we will leave these topics undiscussed, despite their undoubtful importance as the empirical literature on judicial behavior has already shown.

Voting Panels, Composition, and "Peer Effect"

Before, we have seen some accounts of how panel composition at courts affect judges' voting patterns. Social psychologists and behaviorists have long studied the effects of peer pressure in organizations, mostly enterprises, and one would expect – and actually observe – the same happening in public and political organizations, such as courts (and the Congress, etc.).

Epstein, Landes, and Posner (ELP 2013) have a theoretical explanation for the occurrence of panel composition effect, and they test it. Panels

may decide unanimously – when there is no dissenting vote – or non-unanimously – when there *is* dissent. ELP explains that there are costs and benefits of dissenting, and not rarely the former supersede the latter. Costs of dissenting include writing the dissenting opinion, disagreeing with colleagues, and reputational costs infringed on the other members of the panel. All these create *dissent aversion* and, consequently, one will avoid disagreeing on minor issues – specifically technical ones – and dissent more frequently will be caused by ideological disagreements, which are more difficult to resolve by discussion and compromise. ELP (2013) show evidence of this effect for the US Supreme Court from years 1953 to 2008. For other lower courts, the authors predict more dissent when an appeal is being reversed (dissenter has ally in the district court), and less dissent in smaller courts of appeals (judges sit together more frequently, making dissent much costlier). The authors also predict that dissent will be inversely proportional to the court's workload, i.e., the busier judges are, the less they will dissent. Historical evidence from the US Supreme Court and the Courts of Appeals corroborates the authors' predictions. ELP also find that the presence of dissent impact on the length of written voting: majority opinions are longer if there is one dissenting member on the panel, and still longer if there is more than one. Apparently, more words are necessary to justify an opinion when it is faced with opposition. Finally, the authors link the frequency of dissent to one's career: federal judges in the USA tend to dissent more during the first half of their active lives.

Voting panels may also *potentialize* some other factors, for instances, political ideology. As observed by Sunstein et al. (2006, *apud* ELP 2013), judges appointed by Republican Presidents decide against affirmative action cases more often than those appointed by Democratic Presidents; yet, the frequency is much higher for full-Republican panels and much lower for full-Democratic ones. For the gender factor, similar peer effect is observed: all-female panels tend to favor female litigants and the opposite for all-male panels on male litigants (as shown above). However, sometimes, panels may *attenuate* or even *reverse* ideology impact. Studies have found evidence of liberal judges deciding in more conservative manner on panels than they would otherwise do, and the contrary to conservative judges (check on ELP 2013, chapter 2).

Outside the USA, Smyth (2005) studied the pattern of dissent on the Australian High Court. He finds evidence for dissent caused by diverging political ideology, but no evidence of relation between caseload and dissent rate. As to a judges' active career, Smyth finds evidence of *increasing* dissent rate throughout the time, a diverging result to what ELP have shown for the USA.

Related to peer effects, some interesting literature on the influence of social norms on judicial behavior may also be accounted for. Using an analytical approach, Harnay and Marciano (2004) build up a model in which they show that judicial behavior is not entirely a product of individual calculation; instead, it reflects the interactions in a system where judges do care about what other professionals in the community think and do and have desires for certain degree of conformity. In other words, they show evidence that judicial behavior cannot be explained solely by the economic theory from Posner, but also by elements of the strategic theory. This "tendency to conform" to precedents by a particular judge is a result of a cost-benefit analysis: he/she analyzes the expected gains of deviation versus the gains of compliance with the precedents and does that at the private and professional levels. This, according to the authors, explains why judges behave in conformity to precedents under certain circumstances and deviate under others.

External Pressures: Media, Popular Opinion, and Interest Groups

Besides the effect exerted by peers in voting panels as discussed above, there are other sources of external factors that might influence judicial decisions. Media and popular opinion have always restrained, somehow, public agents' behavior; yet, the intensity of this impact has grown exponentially with the modernization of telecommunication technology. In some countries, Supreme Court voting sessions are broadcast live on TV channels. Although average citizens can

rarely comprehend the matters discussed at courts – and especially high courts – due to their complexity and technicality, from times to times Justices' decisions are in the spotlight, highlighted on the front pages of newspapers, in the evening TV shows, and discussed by lay citizens. Thus, even judges who are not directly elected feel, somehow, constrained by what society has to say about the outcomes of their work. As Epstein and Kobylka (1992) pose: "Most modern Court decisions reflect public opinion. When a clear-cut poll majority or plurality exists, over three-fifths of the Court's decisions reflect the polls. By all arguable evidence the modern Supreme Court appears to reflect public opinion as accurately as other policy makers" (p. 24).

Organized interest groups may also exert considerable impact on judicial decisions; although not new – at least in the North-American context – this is a phenomenon that has grown recently, and that is still not well understood (Epstein and Kobylka 1992).

Empirical literature on the effects of media, public opinion, and interest groups on judicial decision is also vast and growing. Due to its higher exposure and its greater impact on the rest of society, studies of this kind have mainly focused on supreme courts. Casillas et al. (2011) find significant influence of public opinion on the US Supreme Court decisions and more impacting in nonsalient cases (because in salient cases, there are high stakes in following legalistic considerations and/or personal ideologies). The authors measure the costs that the Supreme Court incurred by ignoring public opinion in nonsalient cases during the period of 1970s to year 2000. Giles et al. (2008) follow the same direction and, even though they are more cautious about the existence of direct impacts of public opinion on Supreme Court outcomes, they do assert that there is evidence of causality on Justices' voting.

Epstein and Martin (2010) also find evidence that Supreme Court decisions are, to a certain degree, aligned with public opinion. Besides the usual explanation that Justices care about their reputation and society's approval, Epstein and Martin argue that the alliance may occur because Justices *are themselves* part of society and of the public. Thus, in this case, they are actually deciding based on their personal ideologies, and not only as a reflection of external preferences. It would not be easy to empirically separate these two effects, and the authors leave the analysis for future studies.

With regards to interest groups, Collins and Martinek (2010) find evidence of their impact on the probability of success by appellants, but not by appellees, on the US courts of appeals. For the Supreme Court, Collins and Solowiej (2007) also show that interest groups, represented by *amici curiae* briefs presented to the Court, do affect the amount of information that would be important for judges' decisions. A positive outcome the authors find at the US Supreme Court is that it is accessible to a wide variety of interest groups, which reflect a democratic and pluralistic society. However, despite their clear influence, at least on provision of information, the authors were not able to explain how and how much do these groups affect judicial decision-making.

A notable remark about the involvement of interest groups in judicial decisions is made by Epstein and Kobylka (1992). They differentiate this type of activity with those of other political pressure groups: "Unlike the more traditional arenas of group lobbying, . . . [the import of interest groups in the Judiciary] is not so much derived from their numbers but is more a function of the kinds of arguments they present to a Court" (p. 306). These authors believe this the manner by which they influence judges' decision.

Pressure coming from the Executive and the Legislature Powers (i.e., the Congress) also exerts significant impacts on judicial decisions. The interplay between judges and those actors has long been discussed by legal scholars and is a never-ending object of study. Especially in the case of Supreme Courts, due to the Presidential nomination of its Justices, the quest for a better understanding of this relationship is related to the crucial matter of independence of powers.

Lopes and Azevedo (2017) find evidence in this direction for Brazil. Under that system, there are two high courts: STJ (Superior Tribunal de Justiça) – for civil appeals – and STF (Supremo Tribunal Federal) – a constitutional court, which is de facto above STJ. Although Justices at both

courts are formally nominated by the President, STF is more political, with closer links to the Executive and Legislative. Presidents also have more discretion in the nomination of STF Justices. Perhaps because of that, this study finds that STF is significantly more impacted by political influences than STJ.

Garoupa et al. (2013) try to understand these influences at the Spanish Constitutional Court. They conclude that, due to the limitations imposed by civil law systems, impacts in this Court do not strictly follow the patterns predicted by the attitudinal model, specifically on the dichotomy liberal x conservative, or left x right.

As of the relations between courts (especially the Supreme Court) and Congress, Epstein and Kobylka pose that "[they] are hardly random. The political composition of the legislature vis-à-vis that of the Court plays a major role in determining the course of those relations, be they antagonistic or amiable" (p. 24). The fact that Supreme Court Justices must be approved by the Senate also poses a constraint to the first. Yet, since policies created by the Congress, if questioned by unsatisfied litigant groups and individuals from society, may be overruled by the Court, the power factor and impact is not a one-way path. Again, due to the limitations of this chapter, we will leave detailed discussions apart. (Further references may be found on Epstein et al. 2013, chapter 2).

Another interesting type of external pressure may affect judicial behavior, specifically at the context of international courts. Under such regimes, in which no strict hierarchy exists between international and local courts, the problem of compliance by these ones is intensified. Dyevre (2016) designs a theoretical model in which he predicts the circumstances under which domestic judges have incentives to accommodate with the European Court of Justice (ECJ). For this purpose, the author employs the concept of *crisis costs*, or costs for domestic noncompliance. The level of these costs is the determinants of the behavior by local judges. His model shows that, in legal regimes as integrated as the EU, noncompliance (or defiance) makes local courts worse off. Because of the magnitude of such costs, there are incentives for the ECJ and these courts, even strong ones – such as the German Federal Constitutional Court – to seek accommodation through dialogue. For smaller domestic courts, there is no substantial benefit to defy the ECJ generating conflicts. Thus, the extent to which pressure may be exerted upon judicial decision may go beyond national borders.

Indeed, the analysis of international courts, such as the ECJ, renders several possibilities to test hypotheses on judicial behavior. (Other insightful research about judicial decision-making at the ECJ may be found at: Josselin and Marciano 1997; Tridimas and Tridimas 2002; Portuese 2012; among others). In this short chapter, it will not be possible for us to dwell extensively on the multiple works on this theme; yet, we close with one another intriguing work: Vaubel (2009), by analyzing data from 42 countries, shows that the ECJ is indeed biased towards political centralization. Furthermore, and most importantly, this bias is a consequence of independent courts (not the contrary), which are not subordinated to the European Parliament and the European Commission.

Benefit Maximization (or Satisfaction)

Another set of factors that might significantly affect judicial decision-making is the desire that judges have to maximize self-benefits. In Posner's terms, this is the so-called economic theory of judicial behavior. One example of benefit that judges normally try to maximize is the chances to succeed in their career. By observing the Italian Constitutional Court, Melcarne (2017) corroborates the hypothesis that judges' career concerns affect their behavior; the reason is the reputational impacts of their conduct. Furthermore, these concerns are independent to judges' personal characteristics. In line with the discussion in the previous item, Melcarne also observed that judges are sensitive to external pressures, especially interests of the Executive Power.

Schneider (2005) also confirms career incentives in judges' behavior; yet, his observations of the German Labor Courts go further, by analyzing theories of internal labor markets (or tournaments) in the judicial structure. As a combination of career concerns and external pressures

(as discussed in the previous item), the author shows that "[j]udges are likely to decide in a way that conforms to the policies or opinions of those agencies who influence their appointment. These agencies can be higher-level courts, parliaments, and governments" (p. 140). He also claims that these results might be generalized to other jurisdictions, due to the similarity of the structure of German labor courts and that found in many other civil-law countries.

However, some researchers are cautious about viewing judges as benefits maximizers, whatever benefits might be translated to. As Christmann (2014) poses "[m]odels that treat judges as rational maximizing men commonly find a stronger distortion in the resolution of disputes" (p. 411). Because of this skepticism, and with the influential developments of behavioral economics, a new derivation of the benefit-maximization theory on judicial decision-making is emerging. Authors following this line see judges not as benefit-maximizers, but as Tsaoussi and Zervogianni (2010) pose, "satisficers" "who make decisions within real-world constraints, such as imperfect information and uncertainty, cognitive limitations and erroneous information" (p. 333). The concept of "satisficer" comes from Herbert Simon's notion of "bounded rationality": instead of fully rational maximizers – who seek the "best" outcome, satisfices make decisions based merely on the choice of "good enough" outcomes. Under this view, judges' goal "is not optimize but to render opinions that are merely satisfactory" (p. 333), and because of that, they often engage in "improper" behavior, especially under nasty environmental conditions, such as overloaded court dockets and increasingly complex lawsuits.

The theory of judges as satisficers is a promising one, however, still needs to be corroborated by empirical analysis. Hopefully, the ever growing literature on behavioral law and economics will, very soon, bring promising results in this direction.

Above, we have shown several factors that, according to a rich empirical literature, may impact judicial behavior. Due to the limitations of this chapter, the discussion was brief and non-exhaustive. However, some authors argue that there is still a long way to go, in the study of the factor impacting judicial decision. They claim that several qualitative factors have been entirely left out the discussion up to now. For instances, Richards (2017) points to "background variables, such as [judges'] education, prior experience, training, socio-economic background, personality," which could "contribute to a more universal understanding of judicial behavior, crossing the boundaries of both space and time" (p. 558). Indeed, much still needs to be done.

Future Directions

What can one infer from this brief discussion and review of the empirical studies on judicial decisions? What have the studies so far evidenced?

At the beginning, the studies had a more *positive* approach: attempts to capture trends in judicial decisions, sources of personal ideology, and other factors that might impact court outcomes.

Throughout the time, as theories and methodologies developed (although, as seen above, there is still much to be pursued), authors of judicial decisions had increasingly (although sometimes not explicitly) adopted a *normative* perspective. Resulting evidence shows that judges are influenced by factors such as panel composition, material and nonmaterial benefits, pressure groups. Based on that, one might ask how institutions should be better designed to provide the "right" incentives for judges to behave in the "desired" manner, by making decisions that are more democratic, more inclusive, more efficient, . . .? (And each one has his/her own desired outcome for judicial decisions.)

Empirical analysis of judicial decisions has helped to answer these questions, too.

References

Bourreau-Dubois C, Doriat-Duban M, Ray JC (2014) Child support order: how do judges decide without guidelines? Evidence from France. Eur J Law Econ 38(3):431–452

Boyd CL, Epstein L, Martin AD (2010) Untangling the causal effects of sex on judging. Am J Polit Sci 54(2):389–411

Cameron CM, Kornhauser LA (2017) Rational choice attitudinalism? Eur J Law Econ 43:535–554
Casillas CJ, Enns PK, Wohlfarth PC (2011) How public opinion constrains the U.S. Supreme Court. Am J Polit Sci 55(1):74–88
Christmann R (2014) No judge, no job! Court errors and the contingent labor contract. Eur J Law Econ 38(3):409–429
Coase RH (1960) The problem of social cost. J Law Econ 3:1–44
Collins PM, Martinek WL (2010) Friends of the circuits: interest group influence on decision making in the U.S. courts of appeals. Soc Sci Q 91(2):397–414
Collins PM, Solowiej LA (2007) Interest group participation, and conflict competition, the U.S. Supreme Court. Law Soc Inq 32(4):955–984
Dyevre A (2016) Domestic judicial defiance and the authority of international legal regimes. Eur J Law Econ 44:453. https://doi.org/10.1007/s10657-016-9551-2
Epstein L (2016) Some thoughts on the study of judicial behavior. William & Mary Law Review 57(6): 2017–2073
Epstein L, Kobylka JF (1992) The Supreme Court and legal change: abortion and the death penalty. Thornton H. Brooks series in American Law & Society. The University of North Carolina Press, Chapel Hill
Epstein L, Martin AD (2010) Does public opinion influence the Supreme Court? Possibly yes (but We're not sure why). J Constit Law 12(2):263–281
Epstein L, Landes WM, Posner RA (2013) The behavior of federal judges. Harvard University Press, Cambridge
Farhang S, Wawro G (2004) Institutional dynamics on the U.S. court of appeals: minority representation under panel decision making. J Law Econ Org 20(2):299–330
Fon V, Parisi F (2006) Judicial precedents in civil law systems: a dynamic analysis. Int Rev Law Econ 26:519–535
Garoupa N, Gomez-Pomar F, Grembi V (2013) Judging under political pressure: an empirical analysis of constitutional review voting in the Spanish Constitutional Court. J Law Econ Org 29(3):513–534
Giles MW, Blackstone B, Vining RL (2008) The Supreme Court in American democracy: unraveling the linkages between public opinion and judicial decision making. J Polit 70(2):293–306
Grezzana S, Ponczek V (2012) Gender bias at the Brazilian superior labor court. Braz Rev Econom 32(1):73–96
Harnay S, Marciano A (2004) Judicial conformity versus dissidence: an economic analysis of judicial precedent. Int Rev Law Econ 23:405–420
Josselin JM, Marciano A (1997) The paradox of leviathan: how to develop and contain the future European state? Eur J Law Econ 4(1):5–22
King KL, Greening M (2007) Gender justice or just gender? The role of gender in sexual assault decisions at the international criminal tribunal for the Former Yugoslavia. Soc Sci Q 88(5):1049–1071
Lopes F, Azevedo PF (2017) Government appointment discretion and judicial independence: preference and opportunistic effects on Brazilian Courts. Working paper – Insper. Available on https://www.insper.edu.br/working-papers/working-papers-2017/government-appointment-discretion-and-judicial-independence-preference-and-opportunistic-effects-on-brazilian-courts/
Melcarne A (2017) Careerism and judicial behavior. Eur J Law Econ 44:1–24
Peresie JL (2005) Female judges matter: gender and collegial decision-making in the federal appellate courts. Yale Law J 114(7):1759–1790
Portuese A (2012) Law and economics of the European multilingualism. Eur J Law Econ 34(2):279–325
Posner RA (2008) How judges think. Harvard University Press, Cambridge
Pritchett CH (1968) Public law and judicial behavior. J Polit 30:480–509
Richards D (2017) Lee Epstein, William M. Landes, Richard A. Posner: the behavior of federal judges: a theoretical and empirical study of rational choice. Eur J Law Econ 43:555–558
Schneider MR (2005) Judicial career incentives and court performance: an empirical study of the German labour courts of appeal. Eur J Law Econ 20(2): 127–144
Schwartz A, Murchison MJ (2016) Judicial impartiality and independence in divided societies: an empirical analysis of the constitutional court of Bosnia-Herzegovina. Law Soc Rev 50(4):821–855
Sherwood RM (2004) Judicial performance: its economic impact in seven countries. Paper presented in the 8th annual conference da international society for new institutional economics (ISNIE), Tucson. Available on http://www.isnie.org
Smyth R (2005) The role of attitudinal, institutional and environmental factors in explaining variations in the dissent rate on the High Court of Australia. Aust J Polit Sci 40(4):519–540
Sunstein CR, Schkade D, Ellman LM, Sawicki A (2006) Are judges political? An empirical analysis of the federal judiciary. The Brookings Institution, Washington, DC
Tate CN (1983) The methodology of judicial behavior research: a review and critique. Polit Behav 5(1): 51–82
Tridimas G, Tridimas T (2002) The European Court of Justice and the annulment of the tobacco advertisement directive: friend of national sovereignty or foe of public health? Eur J Law Econ 14(2):171–183
Tsaoussi A, Zervogianni E (2010) Judges as satisficers: a law and economics perspective on judicial liability. Eur J Law Econ 29(3):333–357
Vaubel R (2009) Constitutional courts as promoters of political centralization: lessons for the European Court of Justice. Eur J Law Econ 28(3):203–222
Weder B (1995) Legal systems and economic performance: the empirical evidence. In: Rowat M et al (eds) Judicial reform in Latin America and the Caribbean – proceedings of a World Bank Conference. World Bank technical paper number 280. The World Bank, Washington, DC
Yeung L, Azevedo PF (2015) Nem Robin Hood nem King John: testando o viés anti-credor e anti-devedor dos magistrados brasileiros. Econ Anal Law Rev 6(1):1–12

Empirical Law and Economics

▶ Empirical Analysis

Empirical Legal Research

▶ Empirical Analysis

Empirical Legal Science

▶ Empirical Analysis

Empirical Legal Studies

▶ Empirical Analysis

Endowment Effect

Ben Depoorter[1,2] and Sven Hoeppner[2]
[1]Hastings College of the Law, University of California, San Francisco, CA, USA
[2]Center for Advanced Studies in Law and Economics (CASLE), Ghent University Law School, Ghent, Belgium

Abstract

A vast body of experimental studies in psychology and economics finds that individuals tend to value goods more and demand higher prices when they own the goods than they would be willing to pay for the good when they do not already own it. Although research on the endowment effect has been done for more than three decades, it's theory, empirical methodology, results, and implications continue to be topics of intense discussion among economists, lawyers and psychologists. In this entry, we review the theoretical framework and empirical evidence on the endowment effect and highlight some implications for law and economics research.

Definition

The term "endowment effect" describes observations of a gap between the willingness-to-pay for a good not owned already and the willingness-to-sell a good in one's possession. In other words, the term describes the tendency that people value things higher merely because they own them.

Introduction

A major tenet of rational choice theory is that individual preferences are endogenous: although individuals may value things differently, preferences exist independently of situational circumstances. This principle, most commonly referred to as "preference exogeneity" (Korobkin 1998a, p. 611), "source independence" (Issacharoff 1998, p. 1735), and "status irrelevance" (Korobkin 2003, p. 1228), plays a crucial role in the economic analysis of law. For instance, when economists analyze how the legal system should distribute entitlements among competing claimants or how to regulate consensual exchanges of entitlements after their initial allocation, it is generally assumed that legal rules do not influence the value of the entitlement itself.

Behavioral research demonstrates however that the assumption of preference exogeneity does not hold in many circumstances. A vast body of experimental studies in psychology and economics finds that individuals tend to value goods more and demand higher prices when they own the goods than they would be willing to pay for the good when they do not already own it (compare literature reviews and extensive references in: Horowitz and McConnell 2002; Korobkin 2003; Morewedge et al. 2009). The observed gap between the willingness-to-pay (WTP) for a good not owned already, on the one hand, and the willingness-to-sell (WTS) a good in one's possession, on the other hand, was termed

"the endowment effect" in a seminal paper by Richard Thaler (1980, p. 44).

Although research on the endowment effect has been done for more than three decades, its theory, empirical methodology, results, and implications continue to be the topics of intense discussion among economists, lawyers, and psychologists (cf. Plott and Zeiler 2005, 2007, 2011; Arlen and Talley 2008; Isoni et al. 2011; Klass and Zeiler 2013). This entry briefly reviews the theoretical framework and empirical evidence on the endowment effect. Moreover, it highlights a few implications for the economic analysis of law.

Empirical Evidence on the Endowment Effect

Early demonstrations of the endowment effect at work led to most striking results. In their pioneering study, Knetsch and Sinden (1984) randomly divided experimental participants into two groups. One group received a lottery ticket for a $50 cash prize. Participants in the other group received $3 in cash. Then the researchers offered to sell or buy the lottery tickets for $3. A striking majority of the ticket holders, specifically 82%, kept their lottery tickets. This suggests that the WTA of these subjects was greater than $3. Moreover, only 38% of the cash owners chose to buy tickets, which suggests a WTP of less than $3.

In another early study that received much attention, Knetsch (1989) endowed some participants with coffee mugs and later offered each of these participants to trade a large Swiss chocolate bar for a mug. Other participants received a large Swiss chocolate bar and later had the chance to trade the chocolate for one of the same mugs. Surprisingly, not only did merely 11% of the mug owners choose to trade them for the chocolate bar. This would be intuitive if the mugs had a higher value. However, also merely 10% of the other participants, who were endowed with the chocolate bar, were willing to trade it in for a mug. In fact, Knetsch (1989) used a third group to provide a baseline measure. Participants in this group had a choice between a mug and a chocolate bar without any initial endowment. The results suggest that participants perceived the two goods as roughly equivalent on average: 56% chose the mug and 44% opted for the chocolate bar.

While the endowment effect appears to be robust across ages – at least across ages 5 (children in kindergarten) to 20 (college students) (Harbaugh et al. 2001) – the endowment effect varies across different *types of entitlements*. Due to the constraints of the experimental method, endowment effects are mostly observed for small and tangible consumer items [see, however, a study by Dubourg et al. (1994) on endowment effects for small changes in safety features in the purchase of cars]. Meanwhile, no endowment effects appear to occur when involving tokens with a certain value (van Dijk and van Knippenberg 1996). Conversely, this study observed a significant WTA/WTP disparity when the value of the token was uncertain. In line with this finding, the endowment effect is observed to be more pronounced when it is difficult for individuals to compare the two commodities involved (Bar-Hillel and Neter 1996; van Dijk and van Knippenberg 1998). In a similar vein, Shogren et al. (1994) found that the endowment effect is stronger for endowments for which no close substitutes are available. Curiously, endowment effects were absent in experiments involving items with close substitutes (candy bars), although endowment effects appear robust when concerning many other fungible items (coffee mugs, ballpoint pens). A meta-study on the endowment effect finds that endowment effect is strongest for nonmarket goods, next highest for ordinary private goods, and lowest for experiments involving forms of certain value tokens such as money (Horowitz and McConnell 2002). Similarly, the strength of the endowment effect has also been found to vary across various *circumstances*. Findings in an experiment by Loewenstein and Issacharoff (1994) suggest that the endowment effect is stronger when the commodity is obtained as a result of skill or performance as opposed to chance or random luck.

The endowment has also been linked to *status quo bias* (Samuelson and Zeckhauser 1988), i.e., a preference for the current state that biases people against both selling and buying goods. This

explanation finds some support in a series of field experiments by Hartman et al. (1991).

Critics have challenged endowment effect findings as resulting from flawed experimental designs and "strategic heuristics" (i.e., subconscious applications of usually sensible bargaining habits). Plott and Zeiler (2005) claim that the observed WTA/WTP gap disappears when implementing experimental procedures that control for subjects' misconceptions about the value of the traded commodity, for instance, by preventing that participants receive a signal that increases the value of the endowment when it is presented to them as a gift by the experimenter. The underlying generalization that the endowment effect is merely an artifact of experimental methodology is a hotly debated issue. As a response to generalizations from Plott and Zeiler's (2005) study, Isoni et al. (2011) find that the WTA/WTP gap is reduced for mugs, but remains for lotteries even when employing Plott and Zeiler's (2005) method. To date, no credible conclusion can be drawn from the debate (see also Plott and Zeiler 2011; Klass and Zeiler 2013).

It has also been remarked that the observed gap between willingness-to-pay and willingness-to-sell might not occur in market settings that involve experienced traders and learning opportunities (cf. Knez et al. 1985; Coursey et al. 1987; List 2003). For instance, Coursey et al. (1987) reported results that indeed suggest that the WTA/WTP gap may diminish in such settings. [For a critical discussion of these findings, see Knetsch and Sinden 1987.] Kahneman et al. (1991) ran a set of experiments comparing endowment effects with regard to tokens and consumer goods traded in market setting with repeated interactions that provided learning opportunities. While there was no significant endowment effect on the token market, as soon as the commodity was changed to coffee mugs or ballpoint pens, the market did not clear even in repeated trials. On the one hand, the results indicate that there is no similar effect for money of other value tokens as long as the embodied value is certain (similarly van Dijk and van Knippenberg 1996). On the other hand, strong endowment effects emerged when the traded commodity consisted of consumer goods. In both the market for ballpoint pens and the market for coffee mugs, selling prices were about twice as high as buying prices. Specifically for the mugs, the median mug holder was unwilling to sell below $5.25, while the median buyer was unwilling to pay more than $2.25–$2.75. Experiments involving repetition (cf. Shogren et al. 1994) and studies that employed a subject pool of very experienced traders (List 2004) did not reliably reduce the endowment effect and, therefore, further dispel the notion that the WTA/WTP gap merely results from inexperience.

Recent research has further expanded the psychological foundation of the endowment effect and, in general, retains the central idea that potential sellers contemplate a large loss, whereas potential buyers contemplate a small gain (Nayakankuppam and Mishra 2005; Zhang and Fishbach 2005; Johnson et al. 2007).

Explaining the Endowment Effect

To date, the cause of the endowment effect remains a topic of contention. Experiments have ruled out many of the potential explanations derived from traditional consumer theory, including income effects, transaction costs, information problems, and market mechanisms (cf. Brown 2005; also compare, Knetsch and Wong 2009).

Richard Thaler (1980) initially proposed that the endowment effect results from the different mental treatment of out-of-pocket costs and opportunity costs, which he argued to be mentally coded as loss and foregone gains, respectively. Through this underweighting of opportunity costs, he linked the endowment effect to the asymmetry in value that Kahneman and Tversky (1979) describe as *loss aversion*: the observation that the negative value of giving up an object is greater than the value associated with acquiring it. The fact that losses are weighted more strongly than objectively commensurate gains might help explain endowment effect observations since "if a good is evaluated as a loss when it is given up and as a gain when it is acquired, loss aversion will, on average, induce a higher dollar value for owners than for potential buyers, reducing the set of mutually accepted trades" (Kahneman et al. 1990, p. 1328).

To clarify the impact of loss aversion, one study specifically investigated whether the gap and low trading volume resulted from a reluctance to sell or a reluctance to buy. The experiment involved three groups. Only the "sellers" received a coffee mug and were asked whether they would be willing to sell the mug at each of a series of prices between $0.25 and $9.25. The "buyers" were asked whether they would be willing to buy a mug at the same prices. Finally, the "choosers" were asked to choose at each of the prices between receiving a mug or the money. Recognizing that the "sellers" and "choosers" are in objectively identical situations – deciding at each price between the mug and the money – reveals a surprising pattern: the median reservation prices were $7.12 for the "sellers," $3.12 for the "choosers," and $2.87 for the "buyers." The results suggest that the observed patterns are "produced mainly by owner's reluctance to part with their endowment, rather than buyers' unwillingness to part with their cash" (Kahneman et al. 1991, p. 196).

Loss aversion, however, is merely a descriptive construct employed by Kahneman and Tversky (1979) to explain the asymmetry in prospect theory's value function. It begs the further question as to why individuals experience it. Psychologists have stated that loss aversion may stem from a preference to avoid *regret* and/or to remain *consistent* in decision-making. Since giving up an endowment has a higher potential of causing future regret than not obtaining the endowment, loss aversion may be the result of a regret-avoidance strategy (cf. Gilovich and Medvec 1995; Landman 1987). The desire for consistency (cf. Cialdini et al. 1995) comes into play after the endowment is assigned and a sense of entitlement is established. Being confronted with the contradictory idea to sell the endowment might cause cognitive dissonance, i.e., excessive mental stress or discomfort (Festinger 1957). In this regard, when a seller increases the reservation price (lower WTA), this can be regarded as an attempt to reduce the dissonance created by selling something that had just been acquired and associated with oneself.

Martinez et al. (2011), Lin et al. (2006), and Zhang and Fishbach (2005) have identified *emotions* as a moderating variable on the magnitude of the endowment effect. In their studies, when people are induced to positive emotional states rather than negative emotional states, the WTA/WTP disparity is bigger.

Finally, there is some discussion to what extent ownership itself may account for the disparity between the WTP and the WTA or, alternatively, whether tangent explanations such as loss aversion are involved. On the basis of their analysis and experiments, Plott and Zeiler (2005, 2007) argue that ownership in and of itself does not induce a gap between WTA and WTP. However, Morewedge et al. (2009) disentangle ownership from loss aversion in a series of experiments where brokers buy and sell mugs that they do not own. The endowment effect was observed to disappear when buyers were owners and when sellers were not, suggesting that ownership played a role in causing the WTA/WTP difference. One interpretation is that ownership, by way of association and self-identification with the object in possession, may induce endowment effects.

Implications for the Economic Analysis of Law

It has been noted that the endowment effect is "the most significant single finding from behavioral economics for legal analysis to date" (Korobkin 2003, p. 1227). Although the endowment effect is relevant in many areas of law (for a review of the various legal applications of the endowment effect in the law review literature to date, see Klass and Zeiler 2013), the implications for the economic analysis of property law and contract law are especially noteworthy.

Economic Analysis of Property Law: Initial Allocation of Property Rights, Coasian Bargaining, and Remedies

The endowment effect and the associated behavior anomalies of loss aversion and status quo bias have implications for the Coase Theorem. In the presence of endowment effects, the initial assignment of property rights may affect the final outcome, even if transaction costs are insignificant (Korobkin and Ulen 2000). By making a right

holder more reluctant to trade away the item in an efficient transaction, the endowment effect complicates the efficient reallocation of property rights through bargaining. This alters some of the traditional normative implications in the economic analysis of property rights. First, policymakers might consider allocating property rights to the highest-valuing user – even if transaction costs are low. Second, the endowment effect increases the appeal of damage remedies as compared to property rule protection. If parties who litigate contested entitlements and who are awarded injunctive relief by a court are prone to exhibit endowment effects, they may be particularly unlikely to bargain away their court-approved entitlement, even when the opposing party places a higher value on it (cf. Jolls et al. 1998). Indeed, a series of survey experiments by Rachlinsky and Jourden (1998) indicates that injunctive remedies induce an endowment effect by way of ownership. For instance, when considering the protection of an environmental resource, participants were more insistent on protecting the conservation of the resource than was the case when the same environmental resource had liability rule protection. In contrast to injunctive relief, damage remedies allow higher-valuing individuals to take the entitlement and then pay the market price via the remedy (Korobkin and Ulen 2000). Thus damage remedies may circumvent endowment effects that would otherwise complicate an efficient allocation of rights.

Economic Analysis of Contract Law: Default Versus Mandatory Rules, Explicit and Implicit Opt-in and Opt-out Mechanisms, Standard Form Contracts, and Remedies

First and foremost, the status quo bias interpretation of the endowment effect has implications for the regulation of contract default rules. The endowment effect helps explain the observed "stickiness" of legal default. Even if parties would not likely select an undesirable contract provision, they often neglect to remove this provision when it is assigned to them by default rule. Contracting around undesirable defaults appears to be difficult even when transaction costs are low enough to make it efficient to do so (cf. Korobkin 1998a).

Likewise, the legal endowment effect helps explain why individuals find it hard to opt out of defaults (Johnson et al. 1993). In an experiment, two groups of participants had to choose between alternative automobile insurance policies. One group was presented with a default involving a cheaper policy that restricted the right to sue for pain and suffering resulting from a car accident. Participants in this group were offered the option to instead acquire the full right to sue at an 11% increase of the premium (opt-in). Another group was given a default insurance policy that contained no initial restriction on the right to sue. At the same time, however, participants in this group had the option to forego the right to sue in exchange for a reduction of their insurance premium equivalent to the increase of premiums presented to group one (opt-out). Although standard economic reasoning would predict that the choice between the two options should be the same, the applicable default affected the selection of the policy by participants. While only 23% of participants of the first group elected to opt in to acquire the full right to sue, 53% of the second group elected to retain the full right to sue. Moreover, participants with the restricted right default were willing to pay an 8% average increase of the premium to acquire the full right. When the full right was already the default, however, participants indicated an average willingness to pay 32% more for full coverage than for limited coverage. Overall then, the applicable default affected the preferences of participants in both groups. Note in this regard that also mandatory rules – like withdrawal rights in contract law – can lead to implicit opt-in or opt-out scenarios (Hoeppner 2012). Based on these findings, the endowment effect deserves consideration whenever rule-makers contemplate making use of default rules and opt-out or opt-in mechanisms.

Second, terms and provisions embodied in standard form contracts may likewise serve as reference points for contracting parties in negotiating and drafting contracts. If standard contract form an expectation baseline, they may trigger loss aversion. As a result, even when provided the opportunity to bargain for more efficient terms, contract terms may be biased toward the provided standard forms (cf. Kahan and Klausner

1996). A study by Korobkin (1998b) confirms that when default terms and preexisting form terms offered potential conflicting reference points, negotiation outcomes were biased toward the terms provided in the standard form contracts.

Third, endowment effects have been found to affect contract breach and enforcement decision by way of the applicable remedy. In a laboratory experiment, Depoorter and Tontrup (2012) observed that participants were more likely to reject efficient contract breaches and enforce contracts – even if it is against their financial interests – when the legal default is specific performance, as opposed to damages. In other words, possession of one remedy over the other framed the moral intuition such that the performance of the original contract was experienced more severely to contract promises.

Overall, endowment effects suggest that the statutory framework of contract law is likely to affect the substantive content of contracts even if contract law is based on the notion of private autonomy (freedom of contract).

Conclusion

The endowment effect has been heralded as one of the most important findings of behavioral research (Korobkin 2003). As the empirical findings, methodology and theoretical explanation remain a topic of intense discussion among economists and psychologists, the concept has already been widely applied in both law and economics and mainstream legal scholarship (cf. Klass and Zeiler 2013). As the premises and foundations of the endowment effect continue to be explored, the concept will likely find its way in both theoretical and empirical economic analysis of law.

Cross-References

- ▶ Behavioral Law and Economics
- ▶ Bounded Rationality
- ▶ Coase Theorem
- ▶ Consumer Bias
- ▶ Experimental Law and Economics

References

Arlen J, Talley E (2008) Introduction: experimental law and economics. In: Arlen J, Talley E (eds) Experimental law and economics. Edward Elgar, Northampton, pp xv–lxi

Bar-Hillel M, Neter E (1996) Why are people reluctant to exchange lottery tickets? J Pers Soc Psychol 70:17–27

Brown TC (2005) Loss aversion without the endowment effect, and other explanations for the WTA-WTP disparity. J Econ Behav Organ 57:367–379

Cialdini RB, Trost MR, Newsom JT (1995) Preference for consistency: the development of a valid measure and the discovery of surprising behavioral implications. J Pers Soc Psychol 69:318–328

Coursey DL, Hovis JL, Schultze WD (1987) The disparity between willingness to accept and willingness to pay measures of value. Q J Econ 102:670–690

Depoorter B, Tontrup S (2012) How law frames moral intuitions: the expressive effect of specific performance. Ariz Law Rev 54:673–717

Dubourg WR, Jones-Lee MW, Loomes G (1994) Imprecise preferences and the WTP-WTA disparity. J Risk Uncertainty 9:115–133

Festinger L (1957) A theory of cognitive dissonance. Stanford University Press, Stanford

Gilovich T, Medvec VH (1995) The experience of regret: what, when, and why. Psychol Rev 102:379–395

Harbaugh WT, Krause K, Vesterlund L (2001) Are adults better behaved than children? Age, experience, and the endowment effect. Econ Lett 70:175–181

Hartman R, Doane MJ, Woo CK (1991) Consumer rationality and the status quo. Q J Econ 106:141–162

Hoeppner S (2012) The unintended consequence of doorstep consumer protection: surprise, reciprocation, and consistency. Eur J Law Econ 38:247–276

Horowitz JK, McConnell KE (2002) A review of WTA/WTP studies. J Environ Econ Manag 44: 426–447

Isoni A, Loomes G, Sugden R (2011) The willingness to pay-willingness to accept gap, the "endowment effect", subject misconceptions, and experimental procedures for eliciting valuations: comment. Am Econ Rev 101:991–1011

Issacharoff S (1998) Can there be a behavioral law and economics? Vanderbilt Law Rev 91:1729–1745

Johnson EJ, Hershey J, Meszaros J, Kunreuther H (1993) Framing, probability distortions, and insurance decisions. J Risk Uncertainty 7:35–51

Johnson EJ, Haubl G, Keinan A (2007) Aspects of endowment: a query theory account of loss aversion for simple objects. J Exp Psychol Learn Mem Cogn 33: 461–474

Jolls C, Sunstein CR, Thaler RH (1998) A behavioral approach to law and economics. Stanford Law Rev 50:1471–1550

Kahan M, Klausner M (1996) Path dependence in corporate contracting: increasing returns, herd behavior and cognitive biases. Wash Univ Law Q 74:347–366

Kahneman D, Tversky A (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–292
Kahneman D, Knetsch JL, Thaler RH (1990) Experimental tests of the endowment effect and the coase theorem. J Polit Econ 98:1325–1348
Kahneman D, Knetsch JL, Thaler RH (1991) The endowment effect, loss aversion, and status quo bias. J Econ Perspect 5:193–206
Klass G, Zeiler K (2013) Against endowment theory: experimental economics and legal scholarship. UCLA Law Rev 61:2–64
Knetsch JL (1989) Endowment effect and evidence of nonreversible indifference curves. Am Econ Rev 79: 1277–1284
Knetsch JL, Sinden JA (1984) Willingness to pay and compensation demanded: experimental evidence of an unexpected disparity in measures of value. Q J Econ 99:507–521
Knetsch JL, Sinden JA (1987) The persistence of evaluation disparities. Q J Econ 102:691–695
Knetsch JL, Wong WK (2009) The endowment effect and the reference state: evidence and manipulations. J Econ Behav Organ 71:407–413
Knez P, Smith VL, Williams AW (1985) Individual rationality, market rationality, and value estimation. Am Econ Rev 75:397–402
Korobkin RB (1998a) The status quo bias and contract default rules. Cornell Law Rev 83:608–687
Korobkin RB (1998b) Inertia and preference in contract negotiation: the psychological power of default rules and form terms. Vanderbilt Law Rev 51:1583–1651
Korobkin RB (2003) The endowment effect and legal analysis. Northwestern Univ Law Rev 97:1227–1293
Korobkin RB, Ulen TS (2000) Law and behavioral science: removing the rationality assumption from law and economics. Calif Law Rev 88:1051–1144
Landman J (1987) Regret and elation following action and inaction: affective responses to positive versus negative outcomes. Pers Soc Psychol Bull 13:524–536
Lin CH, Chuang SC, Kao DT, Kung CY (2006) The role of emotions in the endowment effect. J Econ Psychol 27:589–597
List JA (2003) Does market experience eliminate market anomalies? Q J Econ 118:47–71
List JA (2004) Neoclassical theory versus prospect theory: evidence from the market place. Econometrica 72: 615–625
Loewenstein GF, Issacharoff S (1994) Source dependence in the evaluation of objects. J Behav Decis Mak 7:157–168
Martinez LF, Zeelenberg M, Rijsman JB (2011) Regret, disappointment and the endowment effect. J Econ Psychol 32:962–968
Morewedge CK, Shu LL, Gilbert DT, Wilson DT (2009) Bad riddance or good rubbish? Ownership and not loss aversion causes the endowment effect. J Exp Soc Psychol 45:947–951
Nayakankuppam D, Mishra H (2005) The endowment effect: rose-tinted and dark-tinted glasses. J Consum Res 32:390–395
Plott CR, Zeiler K (2005) The willingness to pay-willingness to accept gap, the "endowment effect", subject misconceptions, and experimental procedures for eliciting valuations. Am Econ Rev 95:530–545
Plott CR, Zeiler K (2007) Exchange asymmetries incorrectly interpreted as evidence of endowment effect theory and prospect theory? Am Econ Rev 97: 1449–1466
Plott CR, Zeiler K (2011) The willingness to pay-willingness to accept gap, the "endowment effect", subject misconceptions, and experimental procedures for eliciting valuations: reply. Am Econ Rev 101: 1012–1028
Rachlinksy JJ, Jourden F (1998) Remedies and the psychology of ownership. Vanderbilt Law Rev 51: 1541–1582
Samuelson W, Zeckhauser R (1988) Status quo bias in decision making. J Risk Uncertainty 1:7–59
Shogren JF, Shin SY, Hayes DJ, Kliebenstein JB (1994) Resolving differences in willingness to pay and willingness to accept. Am Econ Rev 84:255–270
Thaler RH (1980) Toward a positive theory of consumer choice. J Econ Behav Organ 1:39–60
van Dijk E, van Knippenberg D (1996) Buying and selling exchange goods: loss aversion and the endowment effect. J Econ Psychol 17:517–524
van Dijk E, van Knippenberg D (1998) Trading wine: on the endowment effect, loss aversion, and the comparability of consumer goods. J Econ Psychol 19:485–495
Zhang Y, Fishbach A (2005) The role of anticipated emotions in the endowment effect. J Consum Psychol 15:316–324

Energy Governance: EU-Russia Gas Exchanges

C. Locatelli
GAEL, CNRS, Grenoble, INP, INRA, University of Grenoble-Alpes, Grenoble, France

Definition

Strong and lasting interdependencies exist between EU and Russia concerning natural gas exchanges. They imply some questions about energy security from both parties. One could have expected the setting up of more or less institutionalized governance structures – as an institutional system of rules – allowing the management of risks and externalities linked to this interdependence.

On both sides there is a willingness to cooperate. However, to date, the institutional gap between the supply and demand of the cooperation is a constraint to define a governance structure. The question is to know if international standards based on rules generated by the EU are consistent with Russia's institutional environment. The competitive logic and the regulation on which the EU energy policy is founded conflict with the institutional specificities of the Russian economy. This contradiction explains the failure in the European strategy of external governance with Russia. The "gas conflicts" between the EU and Russia can be viewed as the result of the confrontation of different models that structure the natural gas industries of these two countries.

Through this study case, the question of the internationalization of the rules, norms, and more generally an institutional system of regulation is clearly posed.

Introduction

The determinant factors of the relation between Russia and the EU concerning natural gas exchanges are the subject of an impressive literature. The strong and lasting interdependencies that exist between the two zones imply some questions about energy security from both parties. One could have expected the setting up of more or less institutionalized governance structures allowing the management of risks and externalities linked to this interdependence (Escribano 2015). This was the aim of the energy partnership and its prolongation in a strategic partnership, the Energy Charter – the first constraining multilateral treatise in the field of energy, notably in the matter of investments (Wälde 2008; Kustova 2016). It was an attempt to institutionalize a cooperative zone at the level of EU-Russia to respond to the stakes of energy security. However, their painful implementation bears witness to the constraints that weigh on the definition of this zone of cooperation, notably because of strong divergences concerning the preferences of those involved in this relation. The failures in this domain are an illustration of the problems linked to all attempts to internationalize the norms, rules, and regulations defined in agreement with the national institutional framework.

Interdependences in the Natural Gas Exchanges Between Russia and the EU: Different Preferences and Interests

The interdependences in the natural gas exchanges between the EU and Russia are not questionable. The Russian gas company, Gazprom, is the sole Russian gas exporter to Europe because of its monopoly on exportations by pipelines. It supplies 30% of the EU imports. Since the 1990s, this represents more than 100 Gm^3/year of natural gas exported. At the same time, the characteristics of the gas networks (specific assets, non-re-deployable assets according to Williamson's transaction cost theory 1985) founded the interdependence. Thus, Europe (70.8% of total Russian gas exports) is the first and foremost export market for Russia. Because of the infrastructures in place, in the short term, Russia is dependent on this market and cannot reorient its exports toward other destinations. This dependency is especially important for the State since the revenue linked to hydrocarbon exports (oil and gas) is an essential factor of the budget equilibrium and of economic growth.

In this relation of interdependence, the position of the Russian state (interests and preferences) is defined by its role as a hydrocarbon supplier (Romanova 2014). Its main aim is to have stable markets in terms of the quantities exported and the levels of prices. Therefore, it is to secure access to its main export market, the EU (Kratochvil and Tichy 2013). A double aim is imposed on Gazprom: to maintain its market share in the European gas market (the strategic aim of the State, its principal shareholder) and to maximize its income from exports (Boussena and Locatelli 2016; Yang et al. 2016). The result is a concept of demand security. For the EU, given its position as a major importer (the largest at the worldwide level), the question is one of securing its supply in terms of both price and volume. This defines a

concept of supply security in a specific context, a competitive natural gas market.

Perception of Risks Linked to Interdependence

Since the beginning of the 2000s, the perception of a natural gas security problem has emerged both in the EU and in Russia. It is linked to interdependence between the two zones. According to the analysis proposed by Cherp and Jewell (2014), problems of energy security derive from a price risk and a volume risk.

EU: The Risks Linked to the Gazprom Market Power

In the process of liberalization, taking into account the oligopolistic structure of the gas supply, competitive markets are an important condition of the supply security. One of the primary aims of the EU Energy Policy is to limit the market power of all dominant companies, i.e., those capable of distorting competition. From this point of view, Gazprom, which supplies more than 30% of the external requirement of the EU or over 60% in central Europe and in the Balkan countries, represents a specific risk in the eyes of the EU. This is especially important as more than 50% of Gazprom's shareholding is in the hands of the State. Furthermore, the different gas crises between Russia and Ukraine, and the current conflict, pose the question of reliable transit for Russian gas exports to Europe.

Russia: Uncertainties of the European Gas Market

Since the end of the 2000s, the EU gas market has become more volatile and uncertain, notably for its traditional suppliers. The process of liberalization of the gas industries, the surplus supply, and the weakness of the demand linked to the economic crisis, but also to European policies in favor of renewable energies and the drastic drop in the price of oil, have had serious effects. Firstly these factors all led to the emergence of more competitive gas markets, and secondly they resulted in a drop in the price of natural gas of more than 50% since 2012. These factors are susceptible to put Gazprom's market share in Europe into question. This tendency will continue. In the medium-short term, new LNG suppliers (like the USA, Australia) and uncertainties concerning demand (economic growth, climate policy) mean that the EU is a risky market for Russia and for Gazprom.

These evolutions have led Gazprom to significantly modify its export policy and to adopt strategies that will be different from its traditional contractual framework. Its attachment to long-term take-or-pay (TOP) contracts is known. These contracts have organized its sales to Europe. They have allowed a share in the risks of the price and volumes between the buyer and the seller, notably by guaranteeing the quantities sold annually (removal clause), and guaranteed the development of the gas fields and gas pipelines to Europe. But in the current phase of volatile and uncertain gas markets, Gazprom found itself obliged to reassess this contractual logic. To answer competitive pressures, it had to proceed with more or less major modifications of certain clauses of its long-term TOP contracts (Boussena and Locatelli 2016).

Stakes and Failures of the Governance Structure for Gas Exchanges Between the EU and Russia

The risks linked to this interdependence asked questions about the definition of a governance structure as an institutional system of rules (Lavenex and Schimmelfenning 2009). The primary aim should be the possibility, through cooperation, to limit the extent of conflicts and to stabilize the relations of exchanges between the two zones. On both sides there is a willingness to cooperate. The different attempts to build partnerships, the Energy Charter proposed by the EU or the "Conceptual Approach to the New Legal Framework for Energy Cooperation" developed by A. Medvedev in response to the Charter are some examples. Nevertheless, the cooperation implies some convergence of the actors' preferences. However, to date, the institutional gap

between the supply and demand of the cooperation is a constraint to define a governance structure as demonstrated by the withdrawal of Russia from the process of the ratification of the Energy Charter treaty.

The EU Offer of Cooperation: Building the Markets by the Exporting the Competitive Model

The EU has attempted to manage its relation of dependence with the gas suppliers by the export of its liberal organizational model: governance by competition and EU energy regulation and norms. This form of external governance is an answer to the stakes of security and interdependence through the setting up of a system of common rules and of a high level of institutionalization (Lavenex and Schimmelfenning 2009). In particular, the objective is to overcome the incomplete nature of the competitive reforms concerning the European gas industries. In this way, it is a means to increase the efficiency and the problem-solving capacity of internal liberalization policy (Lavenex 2004). The process of liberalization necessitates opening up upstream of the gas producers and exporting the competitive rules of the EU gas markets. Particularly, the emergence of several gas companies in Russia following the disintegration of Gazprom, potential exporters and therefore in competition on the EU market, would be susceptible to increase the liquidity of the spot markets and render them credible. Next, the aim is to answer the *market failures* that stem notably from an imperfect competition, from the dominant position of certain companies and asymmetric information.

This is the philosophy of the Energy Charter. It is an attempt to build a supranational economic and political integration, to institutionalize an international market system (Andersen and Sitter 2016). Furthermore, it offers guarantees for international investments and allows a competitive principle of nondiscrimination for access to hydrocarbon resources (Haghighi 2007). This access represents a key stake for the "EU gas security" (Boussena and Locatelli 2013). Additionally, the transformation of energy governance of its main suppliers is the essential aim of the EU energy security strategy (Keating 2012). Finally, on the EU market, the gas supply companies behave in compliance with the EU regulation. In this framework, the EU is capable of "normalizing" their behavior. This defines a power of the EU market according to the terminology of Damro (2012) and Goldthau and Siter (2015).

The Institutional Hiatus: Coherence and Complementarities of the EU Rules with the Russian Institutional Environment

The competitive logic and the regulation on which the EU energy policy is founded conflict with the institutional specificities of the Russian economy. This contradiction explains the failure in the European strategy of external governance with Russia. The questions of coherences and complementarities (Aoki 2001; Amable 2016) of organizational models issued from the economic practices of Western markets with the Russian institutional environment and of their transfer as such remains at the heart of analyses of Russian economic reforms (Hausner 1995; Stiglitz 1999; Murell 2001; Roland 2000; Locatelli and Finon 2003). According to this neo-institutionalist approach EU rules must be coherent or complementary with the Russian institutional and economic environment. It is not the case in the gas sector.

The Russian gas market is characterized firstly by the presence of a State company that is vertically integrated with a monopoly of transport and exports (Gazprom). Some competitive fringes exist on larger and larger segments of the Russian market. This dual market depends on dual prices (administrated and free) and on a strict control by the State of the access to hydrocarbon resources. This reform is that which is compatible with the Russian institutional environment. This latter, marked by the weakness of the market institutions (weakness of the *Rule of Law*, ownership rights that remain poorly defined, regressive fiscal system etc.), seems incompatible with the competitive model and de-integrated organization of the EU gas industries (Locatelli 2014).

This path of reform, and the contestation by the Russian state of the EU nominative power

(Godzimirski 2015; Newmann and Posner 2015), results in increasing divergences in the preferences and interests of these two actors. Their divergent approach to the organization of the gas industries is at the origin of major regulation conflicts. In this context, it is increasingly hard to define and to put in place a governance structure that will be susceptible to manage the gas security problem. Through this study case, the question of the internationalization of the rules, norms, and more generally an institutional system of regulation is clearly posed.

Cross-References

- ▶ Conflict of Laws
- ▶ Economic Integration
- ▶ Institutional Complementarity
- ▶ Institutional Economics

References

Amable B (2016) Institutional complementarities in the dynamic comparative analysis capitalism. J Inst Econ 12(1):79–103

Andersen S, Sitter N (2016) The external reach of the EU regulatory state: Norway, Russia and the security of natural gas supplies. In: Peters I (ed) The European Union's foreign policy in comparative perspective beyond the 'actorness and power' debate. Routledge, Contemporary European Studies, Oxford, pp 80–97

Aoki M (2001) Toward a Comparative Institutional Analysis MIT Press, Cambridge; London, 560 p

Boussena S, Locatelli C (2013) Energy institutional and organisational changes in EU and Russia: revisiting gas relations. Energy Policy 55:180–189

Boussena S, Locatelli C (2016) Guerre des prix ou instrumentalisation de l'incertitude sur les prix: quelle stratégie pour un fournisseur dominant sur le marché gazier européen ? Cahier de recherche EDDEN, 1

Cherp A, Jewell J (2014) The concept of energy security: beyond the four As. Energy Policy 75:415–421

Damro C (2012) Market power in Europe. J Eur Publ Policy 19(5):682–699

Escribano G (2015) Fragmented energy governance and the provision of global public goods. Global Pol 6(2):97–106

Godzimirski J (2015) Russia-EU energy relations : from complementarity to distrust? In: Godzimirski J (ed) EU Leadership in energy and environmental governance: global and local challenges and responses. Palgrave Macmillan, London, pp 89–112

Goldtheau A, Sitter N (2015) Soft power with a hard edge: EU policy tools and energy security. Rev Int Polit Econ 22(5):941–965

Haghighi S (2007) Energy security: the external legal relations of the European Union with major oil- and gas-supplying countries. Hart, Oxford

Hausner J (1995) Imperative vs interactive strategy of systemic change in central and Eastern Europe. Rev Int Polit Econ 2:249–266

Keating M (2012) Re-thinking EU Energy Security: The Utility of Global Best practices for Successful Transnational Energy Governance. in Kuzemko C, Belyi A, Goldthau A, Keating M. (eds) Dynamics of Energy Governance in Europe and Russia (Palgrave MacMillan), pp. 86–105

Kratochvil P, Tichy L (2013) EU and Russian discourse on energy relations. Energy Policy 56:391–406

Kustova I (2016) A treaty à la carte? Some reflections on the modernization of the energy charter process. J World Energy Law Bus 9:357–369

Lavanex S (2004) EU external governance in 'wider Europe'. J Eur Publ Policy 11(4):681–700

Lavanex S, Schimmelfenning F (2009) EU rules beyond EU borders: theorizing external governance in European politics. J Eur Public Policy 791–818

Locatelli C (2014) The Russian gas industry: challenges to the 'Gazprom model'? Post Communist Econ 26(1): 53–66

Locatelli C, Finon D (2003) Les limites à l'introduction des institutions de marché dans un secteur de rente. Histoire des représentations du marché: Xe colloque, Grenoble, Association Charles Gide pour l'étude de la pensée économique, 25–27 septembre

Murell P (ed) (2001) Assessing the value of law in transition economies. University of Michigan Press, Ann Arbor

Newmann A, Posner E (2015) Putting the EU in its place: policy strategies and the global regulatory context. J Eur Publ Policy 22(9):1316–1335

Roland G (2000) Transition and economics: politics, markets, and firms. MIT Press, Cambridge

Romanova T (2014) Russian energy in the EU market: bolstered institutions and their effects. Energy Policy 74:44–55

Stiglitz J (1999) Whither reform? Ten years of the transition. In Pleskovic B, Stiglitz J (eds) World Bank annual bank conference in development economics, pp 27–56

Walde T (2008) Renegotiating acquired rights in the oil and gas industries: industry and political cycles meet the rule of law. J World Energy Law Bus 1(1):55–97

Williamson O (1985) The economic institutions of capitalism: firms, markets, relational contracting. Free Press, New York

Yang Z, Zhang R, Zhang Z (2016) An exploration of a strategic competition model for the European Union natural gas market. Energy Econ 57: 236–242

Energy Regulation

Régis Lanneau
Law school, CRDP, Université de Paris Nanterre, Nanterre, France

Definition

Energy is a fundamental input in modern economies. Its regulation is probably one of the most complex since it is related not only to economic and development issues but also to environmental, geopolitical, and social ones. This entry is using law and economics to inquire into energy regulation, to show the rationale of its regulation and the contemporary challenges it has to face.

Introduction

Man is probably the only animal which is using energy not only to maintain his bodily functions at work but, and by far, for his own comfort. At the end of the eighteenth century and before the industrial revolution, the world was consuming around 250 million tons of oil equivalent (toe); nowadays this figure is reaching 13 giga toe; this dramatic increase also occurred through the development of new energy sources which led to great transitions in energy mix from renewable energy which dominated the eighteenth and nineteenth century to the raise of fossil fuels, especially coal at the end of the nineteenth and oil since the mid-twentieth century. Energy is now a fundamental – not to say a central – input in all production, it is necessary to provide most services, and it is required by both consumers and producers for heating, cooling, cooking, lighting, and using machines and household appliances and of course transportation. Modern life – as we know it – is dependent on energy consumption; so is economic growth and development.

Energy regulation is probably one of the most complex to design, implement, and assess. It is of course a sensitive political issue since it will not merely have local consequences (for the energy sector) but also powerful systemic ones (e.g., for the economy or for the environment). Moreover, its scope is broader than most other regulation: it not only focuses on production but also addresses questions regarding transportation, distribution, storage, zoning, and use. It can also be addressed indirectly through regulations concerning mines, buildings, cars, safety standards, working conditions, and, of course, the environment. The body of rules is typically huge, poorly systematized, and often cryptic for outsiders. It is probably where law and economics could help to cut some underbrush: the framework it is providing to understand regulations is especially powerful regarding energy regulation (1). Moreover, the economic approach could also help to address some contemporary issues in energy regulation (2).

Before exploring these two topics, it is noteworthy to mention that the idea of energy regulation in the European Union context is not as broad as what the word refers to in the "common language." Indeed, only natural gas and electricity are concerned, oil benefiting from a special treatment. This could easily be explained using an economic approach of these regulations (infra; 1.1; see also Viscusi et al. 2001), but this restrictive understanding is not helping to stress the fact that energy regulation cannot be addressed energy by energy; it would be better to conceptualize it as an ecosystem in which energy are interlinked.

Why Regulate?

Economic analysis is providing a traditional rationale for regulating: market failures. In the case of energy regulation, it allows for a unifying approach of all regulations which is often missing. The economic rationale is of course not the only one. Political rationales are also powerful considering the central function of energy in modern life.

The Economic Rationale

From an economic point of view, regulations are required if markets are revealing "market failures"

(externalities, market power, imperfect information). This market failure approach to energy regulation reveals itself quite powerful to systematize existing regulations and could explain the "special treatment" reserved to electricity and natural gas. Imperfect information regulation being not fundamental to energy regulation will not be considered here.

Externalities

Producing and consuming energy entail external costs which are not fully considered by rational economic agents either because transaction costs are too high to internalize them through a market or because property rights are not well defined (which of course complexify the analysis to pinpoint "real" externalities). Burning coal, gas, or oil to produce electricity leads to the production of carbon dioxide, sulfur dioxide, and other pollutants; using uranium to fuel a nuclear plant creates radioactive waste. If discharging pollutants (or emitting waste) in the atmosphere is not perceived as a cost, overproduction is to be expected. Even worse, in such system, producers will not have the right incentive to substitute one source of production for another because they will only consider their private costs and not the social cost of their activity.

These externalities are probably the most famous, but many other sources exist:

1. Storage safety (e.g., of oil or gas)
2. Location (e.g., of plants, pipelines, transformer substations)
3. Transmission (because of the physical characteristics of electricity, a coordination is required between producers to power the grid; interconnection of networks also required some coordination)
4. Transportation (because energy fuels are hazardous materials)

Externalities in the energy sector are often addressed through environmental, safety, and zoning regulation. The label should not confuse the lawyer or the economists; these are ways to regulate energy.

Market Power

It is at this level that a "labeled" energy regulation exists. If it is true that market power could also be – and is – addressed ex post through competition law (e.g., through the doctrine of essential facilities), some specific characteristics of gas and electricity were considered to require specific forms of regulations.

First, because of their network structure requiring substantial physical infrastructure, this sector often exhibits the characteristic of "natural monopolies" at least for part of these infrastructures (inefficient to duplicate) and not necessarily for production or distribution. However, and before restructuring policies introduced in the 1990s in Europe, it was quite common to face a fully vertically integrated entity responsible for production, transmission, and distribution. Since the market discipline is not effective for natural monopoly, nothing is incentivizing the monopoly to improve its performance or services (for an alternative view and more development on the traditional approach, see Posner 1999). Because of this situation, nationalization and price regulation were often used with all their practical difficulties (for a detail of price regulation technics, see Decker 2015; see also Viscusi et al. 2001).

Second, with the restructuring policies of the 1990s (directives n^{os} 96/92/CE and 98/30/CE for gas and electricity in Europe), allowed by technology improvements, an independent system operator in charge of maintaining and operating the network typically appeared. This operator, quite often a nonprofit entity, was and still is often supervised by an independent regulatory agency. These policies were supposed to introduce market discipline with the expected benefits of lowering prices, controlling costs of construction and stimulate innovation (Decker 2015). Their success is mixed (for electricity, see Pollitt 2007).

The Political Rationale

From a political point of view, two main rationales are often mentioned. First, it is required to insure some form of energy independence. Second, territorial cohesion and social justice consideration are also playing a key role in energy regulation.

Energy Independence and Security of Energy Supply

Since energy is the lifeblood of economic systems and especially crucial for defense matter (especially transport of troops and dissuasion), geopolitical consideration often entered in energy regulation. Being, like Ukraine, at the mercy of an energy supplier is perceived as a real geopolitical risk; moreover, limiting risks of an economic paralysis due to a price shock or preventing supply disruption are also considered by energy regulations. The development of the nuclear industry in France proceeds from this logic; the General De Gaulle wanted to insure an energy independence of France. This logic also exists at the level of the European Union. Article 194 TFEU especially mentioned the "security of energy supply" (see also Helm 2002). Council directive 2009/119/EC is "imposing an obligation on Member States to maintain minimum stocks of crude oil and/or petroleum products" (typically 90 days of the average daily consumption of a country).

This security of energy supply could lead to specific regulation in hospital and other public services to make sure that the service will not be interrupted.

Territorial Cohesion and Social Justice

Energy being central to modern life, its provision is often considered a matter of public interest. Reliable access to energy in all parts of the territory is considered as an aim in itself, and the regulators could impose some obligations network operator to make sure that basic provision of energy (electricity especially) is ensure and all parts of the country or a regional zone (typically eastern countries in Europe) are connected. In general, these obligations are leading to compensation (which should not provide an advantage or a disadvantage to the addressee of these obligations). This type of regulation typically does not exist for oil because of its physical characteristics (it can be transported and stored) and relatively secondary importance for basic energy needs.

As a basic need, social fares are also quite common for part of the consumption of energy (the one considered as a "need"). Since these social fares result from a political will, providers are compensated for the difference between the social fare and the regular one. Moreover, some regulations are introduced to limit the possibility of suppliers to withdraw access to "poor" families unable to pay their bills. Once again, a compensation should then be provided.

Contemporary Issues in Energy Regulation

Energy regulations are still facing numerous challenges from pushing toward a new energy mix to conciliating development and climate change and adapting infrastructures to new needs. It would also have been possible to mention the difficulties of achieving efficient regulations in a domain where lobbying and rent-seeking are considered as a norm. In this section, it will not be possible to address all these issues. Two important challenges will only be quickly considered: first, insuring a real competition between energy producers, so that energy will be supplied by efficient methods, and second, regulations are now trying to promote energy efficiency and are more and more addressing not only suppliers but also consumers.

The Challenge of Insuring Real Competition Between Energy Producers: A Way to Develop Green Energies?

The challenge of insuring an effective competition does not only exist between producers but also among producers of a certain type of energy. We already noted that, in the domain of electricity (for both retail and generation, Decker 2015), restructuring policies achieved mix results. Ensuring the effectivity of competition law would already be a first step, and regulated prices in this domain should probably be removed following this logic.

However, the most significant challenge is to ensure an effective competition between energy suppliers, a competition based not on private costs but on social costs. Typically, when regulations are not trying to handle negative externalities resulting from some producers and are also not compensating some others for the positive externalities they are generating, prices are no longer

playing their coordination function. Some costlier energies appear cheaper for consumers which will then not adopt optimal behavior. The energy mix transition (from fossil fuel to renewable energies) which is required by climate change is then stifled (OECD 2012). Even worse, according to OECD, support for the production and consumption of fossil fuel amounted to about USD 55–90 billion every year during the period 2005–2011 for OECD countries (OECD 2013). If such a policy could be easily explained through public choice, it appears clearly inefficient from a macroeconomic point of view. Removing these support measure would certainly spur innovation in and consumption of green energy by lowering incentives to produce and use fossil fuels. It would also allow to reduce and improve the efficiency of subsidies to green energy and green technologies (Harris 2006). Note that it would also be required to compute the environmental cost of each and every technology to assess the "real" price which, considering valuation methodology (see Markandya et al. 2002), will lead to lots of contention.

Because of some form of path dependency, it is nevertheless not certain that such radical evolution of regulation is to be expected in the near future or will happen smoothly (Unruh 2002).

Energy Efficiency and Regulation of Its Use

If consumers and producers were entirely rational, they would switch to energy-efficient technology when this technology is also cost effective. If it is not the case, they will not necessarily invest in energy-efficient machinery, appliances, cars, or even buildings. This situation could arise because of implicit discount rate. The consumer will choose an appliance or a car or consider some improvement of his house if and only if she is perceiving it as a good investment. Nevertheless, energy-efficient technologies are often costlier than other technologies, and integrating the private future benefits on energy costs is not sufficient to lead consumer to make such a choice. Indeed, these benefits should be discounted at their present value, and if the implicit discount rate is too high compared to a "normal" discount rate, consumers will not adopt an optimal behavior. Imperfect information could also explain this situation. From this point of view, building codes and fuel efficiency standards could make sense if and only if they are designed adequately. A less paternalist approach would lead to efficiency labeling which is not without its own side effects.

This type of regulation has only indirect effects on energy generation or distribution but could lower the demand for energy and hence the externalities generated (see Mitchell and Woodman 2010). Considering the difficulties to achieve an optimal regulation relative to energy production, this option should not be disregarded.

Conclusion

From a law and economics point of view, energy regulation should be addressed with the idea that what is to be regulated is a full ecosystem and not some sector-specific problems. The market failure approach of regulation provides an interesting starting point to describe and criticize energy regulations. It also enlightens the fact that in this domain of essential complexity, information requirement for an efficient regulation is gigantic. It is nevertheless well accepted that direct subsidies to fossil fuel should be avoided.

Cross-References

- ▶ Efficient Market
- ▶ Energy Governance: EU-Russia Gas Exchanges
- ▶ Market Failure: Analysis
- ▶ Public Goods
- ▶ Public Interest

References

Decker C (2015) Modern economic regulation. Cambridge University Press, Cambridge

Harris J (2006) Environmental and natural resource economics, a contemporary approach. Houghton Mifflin Company, New York

Helm D (2002) Energy policy, security of supply, sustainability and competition. Energy Policy 30:173

Markandya A, Harou P, Bellù L, Cistulli V (2002) Environmental economics for sustainable growth, a handbook for Practitioners. Edward Elgar Publishing, Cheltenham
Mitchell C, Woodman B (2010) Regulation and sustainable energy systems. In: Baldwin R, Cave M, Lodge M (eds) The Oxford handbook of regulation. Oxford University Press, Oxford
OECD (2012) OECD environmental outlook to 2050. OECD Publishing, Paris
OECD (2013) Inventory of estimated budgetary support and tax expenditures for fossil fuels – 2013. OECD Publishing, Paris
Pollitt M (2007) Liberalization and regulation in electricity systems: how can we get the balance right? EPRG working paper, no. 0724. Available online: https://www.repository.cam.ac.uk/handle/1810/194737
Posner R (1999) Natural monopoly and its regulation. CATO institute, Washington, DC
Unruh G (2002) Escaping carbon lock-in. Energy Policy 36:4299
Viscusi K, Vernon J, Harrington J (2001) Economics of regulation and antitrust. MIT Press, Cambridge, MA

Entrepreneurship

Marta Podemska-Mikluch
Faculty of Economics, Gustavus Adolphus College, Saint Peter, MN, USA

Definition

While one could search in vain for a definition of entrepreneurship acceptable to all entrepreneurship scholars, dominant approaches share a number of characteristics. Typically, an entrepreneur is portrayed as an agent of change who operates in a world of incomplete and dispersed knowledge while facing uncertainty about the future. What differs in the various conceptualizations is the mechanism of overcoming uncertainty and knowledge problems. The three leading approaches portray entrepreneurs as dealing with these challenges through either alertness, creativity, or judgment, with new streams of literature adding hypothesis testing and effectuation to the list. While disagreements on the particulars of what entrepreneurship is and what role it plays in economic theory continue, economists now agree on the strong relationship between entrepreneurship and institutions. In particular, it is now well accepted that whether entrepreneurship produces economic growth depends on the institutional context. The other side of this relationship – impact of entrepreneurship on institutions – only begins to receive attention, as does the role of entrepreneurship outside of the conventional market setting.

Introduction

Economists have yet to agree on a single definition of entrepreneurship. Thinkers interested in entrepreneurship originate from diverse perspectives and evoke entrepreneurship to solve distinctive theoretical puzzles. As a result, there is no agreement in the literature on what entrepreneurship is and what role it should play in economic theory. This being the case, anyone interested in entrepreneurship theory should not feel discouraged. The different conceptualizations of entrepreneurship share a number of characteristics. For example, in all cases, entrepreneur is portrayed as an agent of change who operates in a world of incomplete and dispersed knowledge while facing uncertainty about future. It is the mechanism of dealing with uncertainty and knowledge problems that differs in the different conceptualizations of entrepreneurship. The leading approaches portray entrepreneurs as being able to overcome these challenges due to (1) being equipped with alertness to previously unnoticed profit opportunities (Kirzner 1973), (2) being creative and having leadership skills (Schumpeter (1911) 2008, (1943) 2013; Shackle 1972), (3) exercising entrepreneurial judgment (Knight (1921) 2012; Casson 1982; Foss and Klein 2012), (4) experimentation akin to scientific hypothesis testing (Harper 1996), and (5) effectuation (Sarasvathy 2008).

Interest in entrepreneurship theory has been on the rise for the last three decades (Elkjaer 1991; Venkataraman 1997). This is true not only in economics but also in management. For example, in 1987, Entrepreneurship Division was added to the Academy of Management, and now

entrepreneurship has its own code (L26) in the Journal of Economic Literature classification (Foss and Klein 2012, 23). As can be learned from the numerous sources offering an in-depth survey of the history of entrepreneurship thought (Blaug 1998; Praag 1999; Brouwer 2002), the tradition is not new; its origins can be traced all the way back to Richard Cantillon's 1755 distinction between entrepreneur, employee, and landowner with entrepreneur being characterized as carrying the risk (Cantillon (1755) 2015). Cantillon treated the economy as an interconnected whole where coordination and adjustment were secured through competition and entrepreneurship. Even though other thinkers, for example, Jean-Baptiste Say and John Stewart Mill, followed Cantillon's footsteps and also contributed to the theory of entrepreneurship, the concept disappeared from economic theory almost completely by 1870 (Blaug 1998). Undeniably, while the interest in entrepreneurship is now growing, it is yet to be incorporated into mainstream economic theory, and the disagreements on its relevance continue (Foss and Klein 2012, 24).

Despite their status as idiosyncratic outsiders, two twentieth-century thinkers are associated with entrepreneurship in economics and management literatures: Joseph Schumpeter and Israel Kirzner. Equally well known is the difference in their portrayals of entrepreneurship: while Schumpeter sees entrepreneurship as a disruptive force that pulls the economy away from equilibrium, Kirzner views it as a coordinating, equilibrating force. Partially because neither of these theories made significant in ways into mainstream economics, Casson (1982) and Klein and Foss (2012) suggest a third approach, that of Frank Knight. To understand key aspects of each approach, and how they differ, a closer look is warranted.

Entrepreneurship as Creativity and Leadership

Schumpeter's entrepreneur is an innovator, an agent of change, and a disrupter. Entrepreneurship is the force that breaks away from the routine and drives economic change. Schumpeter first introduced the concept of an entrepreneur as an innovator in his groundbreaking *Theory of Economic Development* (Schumpeter (1911) 2008). While Schumpeter was fascinated by Walras and by the general equilibrium theory, he found the static analysis of neoclassical economics to be incapable of explaining economic change. In response, he aspired to offer a process theory akin to that of Marx. Schumpeter described the tendency toward the equilibrium – the key concept in neoclassical economics – as "the circular flow of economic life" where nothing new ever happened. He reasoned that in a theoretical apparatus of general equilibrium, where the circular flow of production and distribution was in balance, there is no room for novelty. Only when entrepreneur is introduced, novelty enters the model. Entrepreneurs break away from the circular flow by making nonroutine choices, by arranging resources into "new combinations." Without entrepreneurship, economy remains in a state of circular flow. So entrepreneurship is necessary in order to give an endogenous account of economic change; it is the driving force of economic development.

In explaining how entrepreneurs innovate, Schumpeter listed five distinct categories of "new combinations": (1) introduction of a new good, this encompasses introducing a good to unfamiliar customers or creating a higher-quality good, (2) creation of a new method of production, (3) the opening of a new market, (4) the capture of new sources of supply, and (5) a new organization of industry (Schumpeter (1911) 2008, 66). Time is the essence in Schumpeter's conceptualization because time turns innovation into routine. The distinction between innovation and routine, between new and old, is the distinguishing factor between entrepreneurs and other economic agents. Once entrepreneurs build up their business and stop carrying out new combinations, they stop being entrepreneurs and turn into managers. In creating new combinations, entrepreneurs perform a function that is separate from that of capitalist, landowner, laborer, manager, or even an inventor. A manager could also be an entrepreneur but not every manager is an entrepreneur and vice versa. In addition, as can be implied from the five

categories of new combinations, Schumpeter recognized that innovation and invention differ: the former requires leadership and will, the latter requires intellect. For economic analysis the one that matters is innovation because "as long as they are not carried into practice, inventions are economically irrelevant" (Schumpeter (1911) 2008, 88).

Leadership, which manifests itself as the willingness to break away from the routine, is a key characteristic of Schumpeter's entrepreneur. It is the entrepreneurial leadership that in Schumpeter's framework provides the creative energy that propels societies in new directions and breaks existing structures. Therefore, to Schumpeter, entrepreneurship is the locus of leadership in a free society. As later developed in "capitalism, socialism, and democracy," leadership and entrepreneurship are the forces behind creative destruction – a force that creates and destroy the structures of the economy in an ongoing process of innovation and change, making it the essence of economic development (Schumpeter (1943) 2013).

Entrepreneurship as Coordination

While Schumpeter's entrepreneur is usually considered a creative and disruptive force, Kirzner's entrepreneur is a middleman, an arbitrager. Entrepreneurship to Kirzner is a force that drives the market toward equilibrium, not away from it (Kirzner 1973). The entrepreneurial function is first and foremost that of securing coordination among market participants. In his most cited book, *Competition and Entrepreneurship*, Kirzner offers a simple definition of the "pure" entrepreneur as observing "the opportunity to sell something at a price higher than that at which he can buy it" (Kirzner 1973, 16).

Kirzner's entrepreneur operates in a Hayekian world of dispersed knowledge, knowledge that is specific to time and place. In such a world, coordination is a puzzle: how is it possible that despite being equipped with different knowledge, individuals are able to find opportunities for mutually beneficial transactions? (Hayek 1937, 1945). Clearly, in the world of perfect knowledge, people would be aware of the full spectrum of options that might be available to them, and there would be no errors for entrepreneurs to capture. Only in the world of incomplete knowledge, some buyers pay prices that are too high, and some sellers accept prices that are too low. By remaining alert to these errors, the entrepreneur observes the differences in prices and profits from reconciling them.

Just like Schumpeter, Kirzner too is critical of mainstream economics and offers a critique of contemporary price theory, in particular the Robinson-Chamberlin model of monopolistic competition (1973). He argues that neoclassical economists are overly committed to the equilibrium framework and pay no attention to the market process, the process through which equilibrium and coordination are generated. While the neoclassical price theory pays no attention to entrepreneurship, Kirzner argues that entrepreneurship is the driving force of the market economy. Along with other economists of the Austrian tradition, Kirzner maintains that the neoclassical approach relies on overly restrictive assumptions. In their view, prices encountered in a market economy are not equilibrium prices, but rather they should be considered emergent exchange ratios that evolve as entrepreneurs seek to discover arbitrage opportunities. Through the process of discovering arbitrage opportunities, entrepreneurs capture profits, and the economy moves toward the equilibrium. In contrast, if the process of adjustment is viewed as automatic, there is no role for the entrepreneur in the model. What is puzzling and interesting to Kirzner is assumed to happen automatically in the price theory framework.

> I will thus argue that the dominant theory, by emphasizing certain features of the market to the exclusion of others, has constructed a mental picture of the market that has virtually left out a number of elements that are of critical importance to a full understanding of its operation. (Kirzner 1973, 4)

While on the surface Kirzner's approach might appear to be completely different from that of Schumpeter, a number of authors view them as complementary (Hébert and Link 1988; Boudreaux 1994; Choi 1995). For example, Boudreaux argues that the two approaches are

complementary in that they both capture the distinct aspects of the competitive market process. According to Boudreaux, if economists adopted a broader notion of equilibrium – one that include quality adjustments and technological and organizational improvements in addition to price adjustments – then Schumpeter's entrepreneurship could also be viewed as an equilibrating mechanism (Boudreaux 1994). In his own reconsideration of the two approaches, Kirzner argues that entrepreneurs are both alert and creative but that only alertness is relevant to the analysis of the market process (Kirzner 1999). Alertness renders entrepreneurship the driving force of the market process. In contrast, boldness, creativity, and self-confidence are psychological traits. While being bold and creative might contribute to entrepreneur's success, these characteristics are not relevant to entrepreneurship's equilibrating function. It is alertness that coordinates market activities, not boldness or creativity. According to Kirzner, Schumpeter's psychological portrayal of the entrepreneur might be an accurate reflection of real-world entrepreneurs. However, these are secondary features that simply give entrepreneurs an edge, whereas alertness is foundational. Moreover, Kirzner reasserts that entrepreneurship must be viewed primarily as coordinating, not disruptive. Without entrepreneurship, it is impossible to explain market coordination.

While disagreement between Schumpeter and Kirzner received significant attention in the literature, Manne notes that a more prominent and far-reaching disagreement occurred between Kirzner and Demsetz (Manne 2014). Demsetz defended price theory and that entrepreneurship is not a separate category of market activity (Demsetz 1970). Subsequently, Tyler Cowen notes that the discussion of entrepreneurship versus optimization is alike to that between philosophy and poetry (Cowen 2003).

Entrepreneurship as Judgment

Foss and Klein abandon the discussion of creativity versus alertness and instead suggest that Knight's conception of entrepreneurship, as judgmental decision-making under uncertainty, provides a better explanation of the entrepreneurial function and can be most smoothly integrated with the economic literature on the firm (Foss and Klein 2012). The authors define judgment as "decisive action about the deployment of economic resources when outcomes cannot be predicted according to known probabilities" (Foss and Klein 2012, 38). They view judgment as different from alertness because alertness tends to be passive, whereas judgment is active. As the authors explain, alertness is the ability to react to existing opportunities, while judgment refers to the creation of new opportunities (Foss and Klein 2012, 39).

They focus on the conceptualization of entrepreneurship as judgment because they see it as a natural complement to the theory of the firm. To capture this relationship, they turn to Knight who linked profit and the firm to the existence of uncertainty (Knight [1921] 2012). Entrepreneurship requires making a judgment about the most uncertain events, and due to this utmost uncertainty, entrepreneurial judgment cannot be sold on the market, but rather becoming an entrepreneur requires starting a firm. Judgment, as considered by Foss and Klein, pertains employment of resources; therefore, entrepreneurial decision-making is ultimately a decision-making about the employment of resources. Interestingly, and to further separate themselves from Kirzner, the authors argue that opportunities for profit do not exist independently, they are not out there, waiting to be discovered. Rather, they need to be created and manifested in action.

Entrepreneurship in Action

Numerous authors have expanded on Knight, Kirzner, and Schumpeter in a manner that allows us to better understand the mechanism of how entrepreneurs actually deal with uncertainty. One of these authors is David Harper, who adopted the Popperian growth of knowledge approach to illuminate entrepreneurial learning (Harper 1996). Harper postulates that market entrepreneurship is akin to scientific hypothesis testing. Similarly to

Foss and Klein, Harper is not persuaded by the notion of passive alertness and believes entrepreneurial opportunities are not discovered but must be created. Due to uncertainty and human fallibility, there is no way of knowing whether the new opportunity is going to be a success or a failure. Therefore, each time entrepreneurs launch a new product or venture, they are testing a hypothesis. If businesses succeed, the hypothesis is unfalsified, and it remains unfalsified as long as there is sufficient interest to sustain the venture. As in science, rejected hypothesis are valuable: they contribute to the growth of knowledge.

Accordingly, Harper defines entrepreneurship "as profit-seeking activity aimed at identifying and solving ill-specified problems in structurally uncertain and complex situations. It involves the discovery and creation of new ends–means frameworks, rather than the allocation of given means in the pursuit of given ends" (Harper 1996, 3). By exploiting similarities between the market process and the process of scientific knowledge formation, Harper sheds light on how entrepreneurs acquire, use, and disseminate new, often tentative and conjectural, knowledge. Following Loasby, Harper notes that the relationship between market process and scientific progress is not a coincidental metaphor but both describe the same process: the growth of knowledge (Loasby 1989).

Harper argues that since scientists and entrepreneurs are problem-solvers, their activities must be described in terms of human ends and purposes. This systematic approach to uncertainty makes Harper's entrepreneur significantly different from Schumpeter's and Kirzner's conceptualizations. While Schumpeter and Kirzner highlighted the nonrational and intuitive aspects of entrepreneurial behavior, Harper argues that rationality and critical methods of error elimination are crucial for acquiring new knowledge.

Saras Sarasvathy offers parallel insights into how entrepreneurs deal with uncertainty. Sarasvathy contributes to the entrepreneurship scholarship a notion of effectuation, which she defines as an inverse of causation. Causal thinking involves selecting means most suitable for the achievement of predetermined ends. Effectuation, in contrast, starts with given means and seeks to create new ends. While economic agents are typically portrayed as employing only causal thinking, Sarasvathy's entrepreneurs combine causal and effectual thinking. They act based on what they know about themselves and others, focus on the control of resources at hand, and on what kind of effects are within their reach. As they generate the alternatives, they simultaneously assess the qualities of different ends. To Sarasvathy, entrepreneurship and effectuation are crucial in opening economics to imagination and novelty, which in turn has significant analytical implications. For example, Sarasvathy argues that while economists have traditionally assumed existence of such artifacts as firms, organizations, or markets, the creation of these facts cannot be explained without effectuation (Sarasvathy 2001, 2008).

Entrepreneurship and Institutions

Given the interest in the relationship between entrepreneurship and economic growth (Aghion and Howitt 1990; Wennekers and Thurik 1999; Blanchflower 2000), impact of institutions on entrepreneurship begun to attract increasing attention. In his now classic paper, Baumol considers the impact of institutional environment on the relative payoffs to different forms of entrepreneurship (Baumol 1990). According to Baumol, not all forms of entrepreneurship are productive. For example, rent seeking, litigation, and warfare are entrepreneurial in nature but have a negative impact on economic growth. When institutional environments renders payoffs to unproductive or destructive activities higher than payoffs to productive entrepreneurial activities, entrepreneurship does not translate into economic growth. In other words, Baumol argues that changes in the extent of productive entrepreneurial spirits are not random but depend on the rules within which entrepreneurial activity takes place. In a similar spirit, Sautet notes that "what is generally missing in countries with lackluster economic performance is not entrepreneurship as such but the right institutional context for entrepreneurship to take place and to be socially beneficial" (Sautet 2005).

It is now well accepted that institutions determine whether entrepreneurship contributes to economic growth. However, as noted by Boettke and Coyne, to simply point out the set of institutions most conducive to productive entrepreneurship is not the same as actually implementing these institutions (Boettke and Coyne 2003). This opens the question of what drives institutional change. While various theories of institutional change can be found in the literature, Wagner places institutions along with market phenomena as an outcome of entrepreneurial action and suggests bidirectional, entangled relationship between entrepreneurship and institutions (Wagner 2009, 2010). In a similar vein, and building upon the examination of institutional entrepreneurship through the lenses of institutional theory and institutional economics (Pacheco et al. 2010), Kuchař uses the emergence of surrogate motherhood market to illustrate how entrepreneurs drive institutional change by challenging existing institutional legal ordering and common interpretations of social phenomena (Kuchař 2016).

Directions for Future Research

If Sarasvathy and Venkataraman are correct in suggesting that we might have miscategorized entrepreneurship as a subfield of management and economics when it is a social force akin to democracy or scientific method (Sarasvathy and Venkataraman 2011), then scholars interested in pursuing entrepreneurship research need not worry about a shortage of research streams. Sarasvathy and Venkataraman go as far as to argue that entrepreneurship is not a subject of social inquiry but a method of human action. There are reasons to believe that they might be correct in anticipating an entrepreneurship-centered revolution, as evidenced by the growing interest in entrepreneurship and the emerging literature on entrepreneurship in unconventional institutional setting, i.e., social, political, institutional, and cultural entrepreneurship (Dacin et al. 2010). Another approach to considering entrepreneurship's role beyond the market setting is that of Podemska-Mikluch and Wagner: entrepreneurial coordination across divergent institutional frameworks, not only within them (Podemska-Mikluch and Wagner 2017). If this broader understanding of entrepreneurship were to take off, one area of economic theory prone to rethinking would be market failure literature because its conclusion relies on the absence of entrepreneurship in any form, be it conventional, institutional, political, or cross-institutional.

References

Aghion P, Howitt P (1990). A model of growth through creative destruction. National Bureau of Economic Research. http://www.nber.org/papers/w3223

Baumol WJ (1990) Entrepreneurship: productive, unproductive, and destructive. J Polit Econ 98(5):893–921

Blanchflower DG (2000) Self-employment in OECD countries. Labour Econ 7(5):471–505. https://doi.org/10.1016/S0927-5371(00)00011-7

Blaug M (1998) Entrepreneurship in the history of economic thought. In: Boettke PJ, Ikeda S (eds) Advances in Austrian economics. JAI Press, Stamford, pp 217–239. https://ideas.repec.org/p/exe/wpaper/9515.html

Boettke PJ, Coyne CJ (2003) Entrepreneurship and development: cause or consequence? Adv Austrian Econ 6:67–87

Boudreaux D (1994) Schumpeter and Kirzner on competition and equilibrium. In: Peter J. Boettke (ed) The market process: essays in contemporary Austrian economics. Elgar, Aldershot

Brouwer MT (2002) Weber, Schumpeter and Knight on entrepreneurship and economic development. J Evol Econ 12(1–2):83–105

Cantillon R ((1755) 2015) Essay on the nature of trade in general. Liberty Fund. https://muse.jhu.edu/book/41671

Casson M (1982) The entrepreneur: an economic theory. Rowman & Littlefield. https://books.google.com/books?hl=en&lr=&id=vo-0aXjiLcoC&oi=fnd&pg=PR10&ots=3-PE11FJXM&sig=G2-othMw9CbJWjZe9zAWx9jNU1g

Choi YB (1995) The entrepreneur: Schumpeter vs. Kirzner. Adv Austrian Econ 2(Part A):55–78

Cowen T (2003) Entrepreneurship, Austrian economics, and the quarrel between philosophy and poetry. Rev Aust Econ 16(1):5–23

Dacin PA, Tina Dacin M, Matear M (2010) Social entrepreneurship: why we don't need a new theory and how we move forward from here. Acad Manag Perspect 24(3):37–57

Demsetz H (1970) The private production of public goods. J Law Econ 13(2):293–306

Elkjaer JR (1991) The entrepreneur in economic theory. An example of the development and influence of a concept. Hist Eur Ideas 13(6):805–815

Foss NJ, Klein PG (2012) Organizing entrepreneurial judgment: a new approach to the firm. Cambridge University Press, Cambridge
Harper DA (1996) Entrepreneurship and the market process: an enquiry into the growth of knowledge. Routledge, London
Hayek FA (1937) Economics and knowledge. Economica 4(13):33–54
Hayek FA (1945) The use of knowledge in society. Am Econ Rev 35(4):519–530
Hébert RF, Link AN (1988) The entrepreneur: mainstream views & radical critiques. Praeger Publishers, New York
Kirzner IM (1973) Competition and entrepreneurship. The University of Chicago Press, Chicago
Kirzner IM (1999) Creativity and/or alertness: a reconsideration of the Schumpeterian entrepreneur. Rev Aust Econ 11(1–2):5–17
Knight FH ((1921) 2012) Risk, uncertainty and profit. Dover Publications, Mineola
Kuchař P (2016) Entrepreneurship and institutional change. J Evol Econ 26(2):349–379. https://doi.org/10.1007/s00191-015-0433-5
Loasby B (1989) The mind and method of the economist. Books. Edward Elgar Publishing. http://econpapers.repec.org/bookchap/elgeebook/288.htm
Manne HG (2014) Resurrecting the ghostly entrepreneur. Rev Aust Econ 27(3):249–258
Pacheco DF, York JG, Dean TJ, Sarasvathy SD (2010) The coevolution of institutional entrepreneurship: a tale of two theories. J Manag 36(4):974–1010
Podemska-Mikluch M, Wagner RE (2017) Economic coordination across divergent institutional frameworks: dissolving a theoretical antinomy. Rev Political Econ. January 29(2):249–266. https://doi.org/10.1080/09538259.2016.1269425
Sarasvathy SD (2001) Causation and effectuation: toward a theoretical shift from economic inevitability to entrepreneurial contingency. Acad Manag Rev 26(2):243–263. https://doi.org/10.5465/AMR.2001.4378020
Sarasvathy SD (2008) Effectuation: elements of entrepreneurial expertise. Edward Elgar Publishing, Cheltenham
Sarasvathy SD, Venkataraman S (2011) Entrepreneurship as method: open questions for an entrepreneurial future. Entrep Theory Pract 35(1):113–135. https://doi.org/10.1111/j.1540-6520.2010.00425.x
Sautet F (2005) The role of institutions in entrepreneurship: implications for development policy. Policy Primer No.1, Mercatus Policy Series
Schumpeter JA ((1911) 2008) The theory of economic development: an inquiry into profits, capital, credit, interest, and the business cycle. Transaction Publishers, New Brunswick
Schumpeter JA ((1943) 2013) Capitalism, socialism, and democracy. Routledge, London.
Shackle GLS (1972) Epistemics and economics: a critique of economic doctrines. Cambridge University Press, Cambridge, UK
van Praag CM (1999) Some classic views on entrepreneurship. De Economist 147(3):311–335. https://doi.org/10.1023/A:1003749128457
Venkataraman S (1997) The distinctive domain of entrepreneurship research. Adv Entrep Firm Emerg Growth 3(1):119–138
Wagner RE (2009) Property, state, and entangled political economy. In: Schafer W, Schneider A, Thomas T (eds) Markets and politics: insights from a political economy perspective. Metropolis, Marburg, pp 37–49
Wagner RE (2010) Mind, society, and human action: time and knowledge in a theory of social economy. Routledge, London
Wennekers S, Thurik R (1999) Linking entrepreneurship and economic growth. Small Bus Econ 13(1):27–56

Environmental Crime

Giuseppe Di Vita
Department of Economics and Business, University of Catania, Catania, Italy

Abstract

To study environmental crime in a perspective of law and economics, it is necessary to identify the protected species from an economic point of view and at the same time to give a legal definition of this kind of criminal behavior. The list of sanctions and their effective deterrence effects in cases of environmental crime are addressed in the final part of this entry.

Definition

Environmental crime is a behavior harmful for the natural environment and its population that is punished with criminal sanctions, according to the nature of the protected species and the type and magnitude of present and future damage.

Environmental Crime

Interdisciplinary issues are hard to approach because they involve different kinds of knowledge with the use of specific terminology; this is particularly true in the case of environmental crime. The analysis moves from the concept of environment as the object of legislative protection and then follows with the definition of

environmental crime. This is not an easy task because even within the legal doctrine there is no unanimous view of the concept of environmental crime, and technical language is not the same for lawyers and economists. The analysis follows with a list of the criminal sanctions that may be applied as a consequence of public enforcement of criminal environmental law against the violators of such rules. In conclusion, the deterrence effects of criminal sanctions are studied in practice, considering the results of empirical studies performed. (See Gray and Shimshack (2011) for a recent survey of the literature on empirical analysis on enforcement of criminal sanctions.)

Environment

From the economic point of view, we should define the interest protected by the law and the reasons why the policy maker (or legislator) chooses to punish illegal behavior against the environment by applying criminal sanctions.

The problem of an optimal use of the environment and its protection originates from its nature of public good (Siebert 1992) and its characteristics of non-excludability and non-rivalry. The inability of the market to determine an equilibrium price may be due to the social costs associated with the private use of the good (as in the case of pollution emissions), or the collective nature of the benefits, that make the excludability cost extremely high and the price market equal to zero (e.g., the case of biodiversity preservation). In the first case, the self-interest of individuals determines an exploitation of the environment, while in the second hypothesis, the indivisibility of the benefits makes the individuals unwilling to pay for this kind of good. Whenever the existence of a source of market failure is present, the market produces an inefficient allocation of resources (Bator 1958), thus rendering necessary the adoption of some economic and legal policy, to bring the system as close as possible to a situation of Pareto-optimal market allocation. A wide spectrum exists of possible measures that it is possible to implement in order to correct the inefficiencies of the market, such as Pigouvian taxes, standards, tradable permits, Coasian solution, and creation of an institution (Faure 2012; Tietenberg and Lewis 2012). This study considers only the cases where the inefficient allocation of environmental assets is punished by means of criminal sanctions.

The main reason for committing a crime against the environment is the saving gained by avoiding a costly compliance to the regulation implemented, adopted with the aim of preserving the environment. The individuals who follow their self-interest, thereby breaking the environmental law, worsen social welfare, overexploiting the environment with irreversible harm for future generations. The economic benefits for violators increase according to the degree of stringency of the legislation (Almer and Goeschl 2015).

From the legislative and court point of view, the kind of remedy is commensurate with the gravity of the behavior to the environment (Rousseau 2009). So the private remedies of repayment for damage or injunction are confined to the less important cases where the consequences of illegal behavior are limited to the parties involved (e.g., the noise in a private apartment). The sources of market failures render the use of standard civil remedies difficult (like using the strict liability regimes, see Sigman 2010) and the administrative fines inadequate.

In the case of private goods, it is sufficient to design the property rights properly in order to bring the system to the optimal path, allowing the victim of damage to file the case in the court and to be reimbursed for the damage, thus indirectly promoting the private enforcement of environmental law. This efficient mechanism cannot be used in the case of public goods because of the incomplete appropriability of the benefits and the multiple offensive environmental damages that go beyond the individual dimension and involve collective values and the welfare of future generations.

To protect environmental goods like air, soil, and water, it is necessary to adopt policies promoting the use of devices to abate the level of pollution and/or to implement more environmentally friendly technologies together with other measures to preserve biodiversity and to avoid climate changes, domestic and trans-boundary pollution, illegal waste disposal in landfill,

pollution emissions, and so on. Often, such kinds of good cannot be efficiently protected by using the standard private law remedies (like injunction to avoid future damages and the condemnation of the author of the offense to pay past damages), in consideration of their characteristic as public good, requiring stronger measures from the policy maker, like a public enforcement of the law, that may only indirectly protect the private rights involved. (About the "civil enforcement tools" of environmental law, see Glicksman and Earnhart 2007.)

Environmental Crime

The "enforcement pyramid" (Ayres and Braithwaite 1995), i.e., the graduation of penalties for illegal behavior that could be applied to protect the environment, is made up of civil sanctions, administrative fines, and criminal sanctions. Although all the kinds of sanctions have as common denominator the protection of the environment by means of punishment of the violator and reimbursement of the damage, the first two kinds of enforcement of environmental law have as their principal aim to deter the violator and compensate the offended individuals, in contrast with the criminal sanctions that have as their main goal to punish the violator.

Among the civil sanctions to implement environmental law, the "citizen's suit" deserves a special mention; it may be defined thus ". . . Citizen suits are court proceedings brought by citizens who seek to enforce public rights. In the environmental arena they are cases brought to enforce the rights or obligations created by environmental laws. They are civil as opposed to criminal cases. While citizens' suits are often brought against government authorities this is not a definitive feature . . ." (Mossop 1993). This measure to implement environmental law has been extensively used in wealthy (Naysnerski and Tietenberg 1992) and less developed countries, precisely since 1958 in the Czech Republic (Earnhart 2000). This kind of law enforcement is based on the citizens' cooperation to protect the environment and may be useful in minor cases of law infringement. (Regarding the effects of alternative policies increasing reliance of self-reporting by polluters, see Farmer 2007.)

The step immediately next to the first in the enforcement pyramid consists of the administrative fines that may be applied without any judge's intervention; their implementation is devoted to public bodies of the state. The problem is that the dissuasive capability of administrative sanctions could be inefficient to punish and deter the more dangerous behavior to the environment. (See on the optimal level of environmental fines the interesting paper by Rousseau and Kjetil 2010.)

In general, it is possible to note that criminal proceedings are, in contrast to civil proceedings, initiated by the government or by public authorities rather than offended individuals. This kind of public enforcement of environmental law requires a high standard of proof, involving fundamental individual rights constitutionally protected (like individual freedom) and moral-social stigma consequent to an application of a criminal sanction.

The environmental crime may be defined from the legal point of view as a crime against the environment or the violation of an environmental law (Clifford and Edwards 1998). (Clifford and Edwards (1998) moreover propose a broad philosophical definition of an "environmental crime" as an act committed with the intent to harm or with a potential to cause harm to ecological and/or biological system and for the purpose of securing business or personal advantage. See Eman et al. (2013) for an updated survey of the definition of "environmental crimes" in general and within the European Union. For a criminological analysis of environmental crime, see Huisman and Van Erp 2013 and Lynch and Stretesky 2014).

Although from the legislative point of view it is simple to say what is an environmental crime, among the international community of law scholars, there is no unanimous definition of this concept; nevertheless, the phrase "environmental crime" is used in several international agreements and national legislation.

Here, we may underline the increasing relevance of the trans-boundary dimension of environmental protection (e.g., air, waste, water, global warming; green crimes include air

pollution, water pollution, deforestation, species decline, the dumping of hazardous waste, etc.). This indicates the need for the international community to define this concept univocally. The main obstacle to this aim is that the identification of environmental goods protected by criminal sanctions depends on the socioeconomic condition of each single country. So you may expect countries at the same stage of development to be more prone to adopt similar or identical legislation and enforcement systems, because their relationship between per capita income and pollution is the same. (About the relationship between per capita income and polluting emissions and the relevance of legal family in explaining the environmental quality levels at different stages of growth, see Di Vita (2008).) Developing countries are not able to use their scarce resources to protect the environment fully. These considerations could be of help in explaining why the delimitation of environmental crime differs among countries.

Criminal sanctions for violation of environmental law constitute a novelty introduced in the 1980s of the last century, with an increasing incidence in recent years (Billiet and Rousseau 2014; Goeschl and Jürgens 2014). Since the 1970s criminal sanctions to protect the environment have become a measure extensively used in the United States and later on in the European Community (Almer and Goeschl 2010). (For an updated survey of "environmental crime" within the European Union, see Öberg (2013).) In the European Union, the criminalization of environmental law enforcement has been strengthened by the European Directive 2008/99/EC (Billiet and Rousseau 2014). More recently, since the mid-1990s, some European countries have introduced administrative sanctions as a complement of criminal fines (Faure and Svatikova 2012; Faure 2012). Administrative measures have the advantage of being more flexible than criminal proceedings so that the enforcement of environmental law may be pursued at lower cost. Even in the former Eastern European countries, the environmental laws are beginning to be protected by the introduction of criminal sanctions (Eman et al. 2013).

Australia is a country where the protection of the environment by means of criminal sanctions has a long tradition and since 1979 has played a leading role in criminal justice processing of environmental offenses through the New South Wales Land and Environment, being probably the first nation to introduce environmental courts (Walters and Westerhuis 2013). Among the so-called BRICS countries, Russia seems to have experienced a reduction of crimes against the environment (according to domestic statistics), although some people suggest that together with a decrease in ecological crimes and offenses, there is some increase of latency within the field of ecological crime. Recent research suggests that one of the major profiles of environmental crime counteraction could be the nongovernmental ecological control system development adopted in Russia that is devolved to both municipal and social ecological control. This raises some doubt about the degree of enforcement necessary to achieve the optimal level of environmental protection (Anisimov et al. 2013).

Even China, one of the most dynamic economies among the developing countries, has recently levied criminal sanctions to punish violators and deter future infringements of environmental law (Zong and Liao 2012). In the African continent, protection of the environment from criminals is not so widespread for the economic considerations discussed previously. Nevertheless, this may represent a missed opportunity to preserve one of the largest green lungs of the world and an inexhaustible source of biodiversity.

The key point of view to consider the importance of environmental crime is the dynamic effects of this kind of sanction in bringing the economic systems toward a more sustainable path to preserve the environment and render the stock of environment useful for future generations.

Criminal Sanctions To Punish and Prevent Behavior Harmful to the Environment

Based on the previous considerations, it is clear that the state has to choose whether to punish some illegal behavior with a criminal sanction in

consideration of the interest protected and the offensive magnitude of present and future damages. In the case of civil remedies to enforce the environmental law, the costs burden the citizens, with limited expenses for the public sector of the economy (state administration and its ramifications). The criminal sanctions are applied because it is assumed that they raise the level of environmental protection, with high costs to the public sector of economy in terms of investigation and imposition of penalties. The enforcement costs are relevant under binding budget constraint.

Shavell (1985, 1987) has indentified five conditions that should be fulfilled to justify the application of nonmonetary sanctions for efficient criminal deterrence. Such conditions are (i) the probability of insolvency of the violator, (ii) the possibility that the criminal is able to avoid paying the monetary sanction, (iii) the illicit earnings for the private agent in infringing the environmental criminal law, (iv) the potential future damage to the environment, and (v) the magnitude of the damage. His conclusion is that the criminal sanctions are worth applying only when these five conditions are strong enough to outweigh the social costs to deter, punish, and hold in jail the authors of environmental crimes (Billiet and Rousseau 2014).

In a general theoretical framework, Emons (2003, 2007) found evidence that criminal sanctions should be applied in the case of repeat offenders, punishing the less harmful violation of the law with monetary penalties; escalating penalties are used not to make the criminal career less attractive but to make being honest more attractive. Despite these theoretical findings, the empirical question is whether the use of escalating sanctions based on offense history is valid when the benefit of crime and the probability of apprehension are high.

The recent guidelines of US sentences suggest inflicting a short period of imprisonment when the environmental crime leads to severe damage (e.g., an illegal dump of waste, river pollution, etc.). In accordance with previous theoretical findings, O'Hear (2004) reports that in Australia, the prison sentences for environmental crime have been limited to the most serious offenses to the environment.

The sanction for a criminal defendant varies according to the crime and includes such measures as the sanction of death, not applicable in cases of environmental law infringement. Incarceration is a period of time that the violator spends in jail, as a consequence of a criminal sentence. Probation is the period during which a person, "the probationer," is subject to critical examination and evaluation. It is a trial period that must be completed before a person receives greater benefits or freedom. (The probationer criminal sanction could be considered quite similar to the suspended sentences that have a lower immediate effect but tend to deter future criminal behavior (Billiet and Rousseau 2014).) Community service is the period of time that the convicted spends doing civil services instead of jail or at the end of the period of imprisonment. Finally, monetary fines are monetary sanctions applied by the courts as a conclusion of a criminal proceeding (for a more complete explanation of the characteristics of criminal sanctions, see Öberg 2013).

Deterrence Effects of Criminal Sanctions

In the introduction we assume that the criminal sanctions are applied in cases of extremely offensive behavior toward the environment. The underground message is that criminal sanctions possess the greatest deterrence capacity to dissuade further environmental damage. The axiom that criminalization raises the level of environmental quality has been questioned several times (Stigler 1970; Polinsky 1980; Andreoni 1991; Heyes 1996), in consideration of greater costs for the use of criminal sanctions and the negative substitute effects on civil and administrative fines.

Recently, Goeschl and Jürgens (2014), moving from the initiatives of the European Commission to prevent environmental crimes, have shown that an increase in the criminalization level does not necessarily lead to higher environmental quality. The greater the area of criminal responsibility for environmental law violation, the lower will be the enforcement by using alternative measures.

Based on these considerations, the debate is shifted into the arena of applied studies.

Cherry's (2001) empirical analysis found strong empirical evidence that financial penalties have a deterrence effect stronger than prison sentences and that monetary sanctions imply a lower social cost to punish those responsible for crimes and to prevent future damages. The limitation of Cherry's study is that crimes against the environment are not considered but illegal behavior in general.

Faure and Heine (2005), in their research on the enforcement of environmental crimes in Europe, find that monetary sanctions are more widespread than criminal sanctions in the Old Continent. Almer and Goeschl (2010) found a positive relationship between environmental crime prosecution and the deterrence effect. This result is reinforced by the outcomes of another empirical study conducted, using microeconomic data, in Germany regarding in particular waste crimes, where persecutors faced inelasticity of 4% of resources devoted to environmental offices in response to a unitary increase in environmental crimes (Almer and Goeschl 2011).

Faure and Svatikova (2012) performed an empirical analysis on the effects of criminal sanction deterrence of illegal behavior against the environment when the penalty is used together with the administrative process to enforce environmental law. Their analysis is based on the theoretical assumption that to enforce environmental law, having a budget constraint, it is more efficient to use administrative fines, instead of criminal sanctions, because they are more flexible and cheaper than criminal proceedings.

Their evidence, based on a dataset available for four Western Europe nations (the Flemish Region, Germany, the Netherlands, and the United Kingdom), shows that administrative fines are a cheaper complement of criminal sanctions.

Billiet and Rousseau (2014) suggest that the relationship between monetary and nonmonetary sanctions is in complementary and not alternative terms. (For an empirical analysis performed on microdata regarding penalties applied by courts in Europe, see Billiet et al. 2014.) In their empirical research, they found that the fines are usually applied together with other criminal sanctions, thus confirming the complementary nature of the different kinds of criminal sanctions.

Billiet and Rousseau (2014) investigated how prison sentences for environmental crime effects are used in fact and their deterrence effects to violators in Flanders, using statistics of seven courts of first instance and the Court of Appeal of Ghent, regarding the complete case law from 2003 to 2007. This empirical analysis confirms that in the EU, countries like Belgium, in 87.49% of criminal proceedings, only a fine sanction was applied to individual violators. These scholars in their interesting study found that the kind of environmental problem does not matter (Billiet and Rousseau 2014) and that prison sentences are not always executed. (Rousseau (2009) performed an interesting empirical analysis of the court sentences in proceedings regarding environmental crime, finding that judges protect the private assets and calculate the sanctions in consideration of the magnitude of the harm).

Almer and Goeschl (2015) performed an empirical analysis on a special kind of environmental crime regarding waste, considering a panel dataset of 44 counties from the German state of Baden-Württemberg between 1995 and 2005. Their outcomes support the theoretical assumption that criminal sanctions have a greater deterrence effect, although the magnitude of this effect depends on the intensity of enforcement. Moreover, the economic conditions and policy measures adopted at jurisdiction level play a significant role to explain why similar law infringements seem to be evaluated and punished in different ways.

The deterrence effect of criminal sanctions to punish the violators of environmental law seems to be a promising field of future research to draw guidelines for further legislative measures.

Cross-References

▶ Commons, Anticommons, and Semicommons
▶ Externalities

References

Almer C, Goeschl T (2010) Environmental crime and punishment: empirical evidence from the German penal code. Land Econ 86(4):707–726

Almer C, Goeschl T (2011) The political economy of the environmental criminal justice system: a production function approach. Public Choice 148:611–630

Almer T, Goeschl T (2015) The Sopranos redux: the empirical economics of waste crime. Reg Stud (forthcoming) 49(11):1908–1921

Andreoni J (1991) Reasonable doubt and the optimal magnitude of fines: should the penalty fit the crime? RAND J Econ 22(3):385–395

Anisimov AP, Alekseva AP, Melihov AI (2013) Current issues of ecological crime counteraction in Russia. Criminol J Baikal Natl Univ Econ Law 3:80–89

Ayres I, Braithwaite J (1995) Responsive regulation: transcending the deregulation debate. Oxford University Press, Oxford

Bator FM (1958) The anatomy of market failure. Q J Econ 72(3):351–379

Billiet CM, Rousseau S (2014) How real is the threat of imprisonment for environmental crime? Eur J Law Econ 37(2):183–198

Billiet CM, Blondiau T, Rousseau S (2014) Punishing environmental crimes: an empirical study from lower courts to the court of appeal. Regul Gov 8(4):472–496

Cherry TD (2001) Financial penalties as an alternative criminal sanction: evidence from panel data. Atl Econ J 29(4):450–458

Clifford M, Edwards TD (1998) Defining 'environmental crime'. In: Clifford M (ed) Environmental crime: enforcement, policy, and social responsibility. Aspen Publishers, Gaithersburg, pp 121–145

Di Vita G (2008) Differences in pollution levels among civil law countries: a possible interpretation. Ener Pol 36(10):3774–3786

Earnhart D (2000) Environmental "citizen suits" in the Czech Republic. Eur J Law Econ 10(1):43–68

Eman K, Meško G, Dobovšek B, Sotlar A (2013) Environmental crime and green criminology in South Eastern Europe – practice and research. Crime Law Soc Change 59:341–358

Emons W (2003) A note on the optimal punishment for repeat offenders. Int Rev Law Econ 23:253–259

Emons W (2007) Escalating penalties for repeat offenders. Int Rev Law Econ 27:170–178

Farmer A (2007) Handbook of environmental protection and enforcement – principles and practice. Earthscan, London/Sterling City

Faure MG (2012) Effectiveness of environmental law. What does the evidence tell us? William Mary Environ Law Policy Rev 36(2):293–338

Faure M, Heine G (2005) Criminal enforcement of environmental law in the European Union. Kluwer Law International, London

Faure MG, Svatikova K (2012) Criminal or administrative law to protect the environment? Evidence from Western Europe. J Environ Law 24(2):253–286

Glicksman R, Earnhart D (2007) Depiction of the regulator-regulated entity relationship in the chemical industry: deterrence-based vs. cooperative enforcement. William Mary Environ Law Policy Rev 31(3):603–660

Goeschl T, Jürgens O (2014) Criminalizing environmental offences: when the prosecutor's helping hand hurts. Eur J Law Econ 37(2):199–219

Gray WB, Shimshack JP (2011) The effectiveness of environmental monitoring and enforcement: a review of the empirical evidence. Rev Environ Econ Policy 5(1):3–24

Heyes AG (1996) Cutting environmental penalties to protect the environment. J Public Econ 60(2):251–265

Huisman W, Van Erp J (2013) Opportunities for environmental crime: a test of situational crime prevention theory. Br J Criminol 53(6):1178–1200

Lynch MJ, Stretesky PB (2014) Exploring green criminology: toward a green criminological revolution. Ashgate, Farnham

Mossop D (1993) Citizen suits – tools for improving compliance with environmental law. In: Cunningham N, Norberry J, McKillop S (eds) Environmental crime: proceedings of a conference held 1–3 September 1993. Australian Institute of Criminology, Canberra

Naysnerski W, Tietenberg T (1992) Private enforcement of federal environmental law. Land Econ 68(1):24–48

O'Hear MM (2004) Sentencing the green-collar offender: punishment, culpability, and environmental crime. J Crim Law Criminol 95(1):133–276

Öberg J (2013) The definition of criminal sanctions in the EU. Eur Crim Law Rev 3:273

Polinsky AM (1980) Private versus public enforcement of fines. J Leg Stud 9(1):105–127

Rousseau S (2009) Empirical analysis of sanctions for environmental offenses. Int Rev Environ Resour Econ 3(3):161–194

Rousseau S, Kjetil T (2010) On the existence of optimal fine for environmental fines. Int Rev Law Econ 30(4):329–337

Shavell S (1985) Criminal law and the optimal use of nonmonetary sanctions as a deterrent. Columbia Law Rev 85:1232–1262

Shavell S (1987) The optimal use of nonmonetary sanctions as a deterrent. Am Econ Rev 77(4):584–592

Siebert H (1992) Environmental quality as a public good. In: Economics of the environment: theory and policy. Springer, Berlin/Heidelberg

Sigman HA (2010) Environmental liability and redevelopment of old industrial land. J Law Econ 53:289–306

Stigler GJ (1970) The optimum enforcement of laws. J Polit Econ 78(3):526–536

Tietenberg T, Lewis L (2012) Environmental and natural resource economics, 9th edn. Pearson Prentice Hall, Upper Saddle River

Walters R, Westerhuis DS (2013) Green crime and the role of environmental courts. Crime Law Soc Change 59(3):279–290

Zong X, Liao G (2012) Sustainable development and the investigation of criminal liability for Chinese environmental pollution. In: World Automation Congress (WAC) proceedings 2012, Puerto Vallarta

Environmental Policy: Choice

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

Abstract

The choice between environmental policies is traditionally considered in terms of ex ante versus ex post interventions on the behavior of (potential) injurers that (can) cause an environmental accident with a consequent environmental damage. Moreover, market-based policies can be implemented as an indirect form of incentive for correct behavior. The two market-based policies, taxes and tradable permits system, are compared on the basis of the difference between price and quantity instruments. And finally, the real situation of the presence of an environmental policy mix is considered as a research challenging topic.

Definition

An environmental policy, on an economic analysis of law point of view, is an instrument to correct malfunctions and subsequent inefficiencies that originate from the presence of market failures: the environment appears as a "public good" that may not be appropriated by anyone and has no market price; the damage to the environment is a case of "externality," in that it is, fully or partly, a social cost that is not internalized into the accounts of the parties causing it (Cropper and Oates 1992).

Introduction

Different policies can be implemented to reach a given set of environmental protection objectives. But nearly all environmental policies consist of two components: the identification of an overall goal (such as a certain level of air quality or an upper limit on emission rates) and the choice of instruments to achieve that goal. In practice, these two components are often linked within the political process, because both the selection of a goal and the mechanism for achieving that goal have important political implications.

"But looking at this problem from a law and economics perspective, we can move from the theoretical definition of the efficiency of different instruments to their practical, and so direct, potential to achieve concrete objectives. In particular, three objectives emerge as relevant in judging the practical efficiency of environmental policies: the first is paying accident compensation to the victims; the second is prevention, in the sense of providing incentives for firms to improve safety standards; and the third is connected with technological change in the sense of encouraging firms to adopt lower-risk technologies" (Porrini 2005, pp. 350, 351).

Hereafter, the choice between environmental policies is considered in terms of ex ante versus ex post interventions on the behavior of (potential) injurers that (can) cause an environmental accident with a consequent environmental damage. Moreover, market-based policies can be implemented as an indirect form of incentive for correct behavior. The two market-based policies, taxes and tradable permits system, are compared on the basis of the difference between price and quantity instruments. And finally, the real situation of the presence of an environmental policy mix is considered as a research challenging topic.

Ex Ante Command-and-Control Policy

Generally, an ex ante policy focuses on the potential injurer's activity from which the damage originates, with effects that precede the occurrence of the same. In the environmental sector, in particular, the standard-setting instrument is most common and consists in the enforcement, by an agency, of a given prevention level in any way that may be quantitatively defined. Practically, this ex ante policy is founded on a centralized structure in charge of setting standards and then ensuring their compliance, according to the

so-called classical "command-and-control" process.

As to the US experience with ex ante policy, the activity of the EPA (Environmental Protection Agency) provides a clear example of policy of this kind implemented by an independent environmental authority. This agency performs its tasks through the enforcement of polluting emission thresholds and the performance of inspections and, possibly, of actions brought to the federal courts.

The choice to develop this ex ante policy provides the advantage of centralized agencies to assure a cost-effective calculation on the basis of the expected damage and of the marginal cost of the different technical preventive instruments. So, following the traditional economic analysis of law approach, well-defined standards generate the correct incentive for the firm to act with caution and take the best production and prevention decisions (Calabresi 1970).

Moreover, in its application, the centralized search facilities, the continual oversight of problems, and a range of regulatory tools make this kind of policy capable of systematically assessing environmental risks and implementing a comprehensive set of instruments. But, on the other hand, as disadvantages, the agencies may not be very flexible in adapting to changing conditions, and a centralized command structure, relying on expert advice, may be subject to political pressure as well as to collusion and capture by the regulated firms.

Summarizing, the choice to implement an ex ante command-and-control kind of policy responds to the need of "uniformity versus flexibility." "To be efficient, such regulations, require information on alternative techniques and a balancing between the profit to the industry and the environmental impacts of various production techniques and processes. Specific production standards may, therefore, rapidly become obsolete" (Faure and Skogh 2003, p. 198).

Ex Post Liability Policy

The most common ex post policy consists in a liability system that provides for a legal authority to identify a party responsible for the damage caused by an environmental accident.

Again, the experience in USA can be considered as an example given that the problem of environmental liability in that country has emerged more than 30 years ago. In fact, in 1980 the Congress issued the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) with the main purpose to bring quick relief and remedy action after an accident, to cope with the "decontamination" of polluted sites, and to recover the cleanup and compensation costs from the liable parties.

The ex post policy is based on the system of liability assignment to the polluting company according to the so-called polluter pays principle. Such principle has been defined as "economic" in that it provides for the economic option of charging the cost of the environmental damage to a specific party that is liable for the accident, rather than generally to the society.

A system of liability satisfies the need to compensate affected parties for physical and economic damage originated from a real event loss and to stimulate preventive measures against future accident.

On an economic analysis of law point of view, in a world of perfect (complete) information, this ex post policy is an efficient method to solve the problem of internalizing the economic effects of environmental accident. In fact, the firms face the proper incentive to take the optimal level of precaution, and the individuals, harmed by pollution, receive the proper compensation, possibly through an insurance provider.

But in practice, the assignment of individual responsibility shows relevant problems that we can summarize in:

(i) a specific polluter could in many cases be difficult to identify because some consequences (i.e., a disease or a reduction in health) may be attributed to a number of different factors besides the pollution and, even if a link between a pollutant and the consequences may be established, it could turn out to be difficult to determine the firm responsible for the damage;

(ii) compulsory insurance contracts that the firms are induced or forced to buy could be incomplete or insufficient because of the difficulty to determine the probability of accident and the distribution of the loss caused by environmental damage, hence making the pricing of the contracts complex;
(iii) the polluter can in some cases be insolvent and unable to pay for cleanup or compensation costs because of the (small) size of the firms operating in dangerous activities in comparison with the (high) costs and penalties of environmental damage.

A liability system can be applied using either a negligence or a strict liability regime. The law and economic literature compares the two regimes considering how they provide a potential polluter with incentives to take adequate preventive measures (Cooter and Ulen 1988). And, generally, strict liability is preferred in the presence of informational issues (Epstein 1973). This could be the case of environmental accidents, where it is particularly difficult to determine the standard to assign liability on the basis of negligence: in reality, pollution has many sources and (potentially) many victims, and it is a hard task to prescribe efficient pollution standards based on a calculus of the abatement cost and the external harm of every source of pollution.

Ex Ante Versus Ex Post Policies

The traditional economic analysis of law approach compares ex ante and ex post policies on the basis of their role in achieving the efficiency goal to stimulate the socially optimal level of preventive care and in so doing controlling the environmental risk.

The authorities responsible for meeting these objectives are the regulatory agencies, which fix standards and check their compliance, and the courts, which assign liability. "Statutory regulation, unlike tort law, uses agency officials to decide individual cases instead of judges and juries; resolves some generic issues in rulemakings not linked to individual cases, uses nonjudicialized procedures to evaluate technocratic information, affects behavior ex ante without waiting for harm to occur, and minimizes the inconsistent and unequal coverage arising from individual adjudication. In short, the differences involve who decides, at what time, with what information, under what procedures, and with what scope" (Rose-Ackerman 1991).

A seminal contribution by Shavell (1984) presents four determinants on the basis of which comparing the two different policy systems.

The first determinant is the difference in information between private parties and the regulatory authority. It clearly could happen that the nature of the activities carried out by the firms is such that the private parties have better knowledge about the cost of preventive care and of the risks involved. In such a case, a liability system is more efficient because it makes the private parties the residual claimants of the control of risks. But it may also happen that the regulator has better knowledge because of the possibility of centralizing information and decisions, in particular when knowledge of risks requires special replicable and reusable expertise. In such a case, direct regulation is likely to be more efficient.

A second determinant is the limited capacity of private parties to pay the full costs of an accident, either because of limited liability or of insufficient assets; this is the case called "judgment proof" (Shavell 1986). An ex post liability policy allows to reach a social optimal solution, only if the private parties can finally cover the damage in the dimension decided by the court. In all the cases in which the damage is superior to the resources of the private parties, a liability system does not provide private parties with proper incentives for preventive care. So the greater the probability or the severity of an accident are and the smaller the assets of the firm are (relative to the potential damages), then the greater the efficiency of ex ante regulation.

The third determinant is the probability with which the responsible parties would face a legal suit for harm caused. This problem is particularly present in environmental risks: in many cases the victims are widely dispersed with none of them motivated to initiate a legal action, harm may appear only after a long delay, and specifically responsible polluters may be difficult to identify.

In the comparison with ex ante policy, ex post liability is more uncertain and so less efficient.

The fourth determinant is connected with the general level of administrative expenses incurred by the private parties and the public. The cost of a liability system includes the cost of legal procedure and the public spending for maintaining legal institutions. The cost of an ex ante policy includes the private costs of compliance and the public spending for maintaining the regulatory agencies. In this case the advantage of the liability system is that legal costs are sustained only if a suit occurs, and, if the system works well, stimulating the efficient level of prevention, the number of suits will decrease, and therefore the costs are becoming lower in time. On the other hand, under an ex ante policy, the administrative costs are sustained whether or not the accident occurs because the process of regulation is costly by itself, and the regulator needs in any case to collect continually information about the parties, their activities, and the risks.

Summarizing, ex ante policy is better when harm is large, is spread among many victims or, takes a long time to show up, when accidents are not very rare events, and when standards or requirements are easy to find and control.

In another contribution, Kolstad et al. (1990) support the hypothesis of the complementarity between ex ante and ex post policies because, even if the economic literature has mainly studied separately the two systems, characterizing each of them by different inefficiencies, the phenomenon of joint use of ex ante and ex post systems is widespread.

Among the determinants presented by Shavell (1984) and above analyzed on the basis of which comparing the two different regulatory systems, the authors concentrate on the fact that potential injurer is in many circumstances uncertain about whether a court will hold him liable in the event of accident and subsequent suit. But an ex ante regulatory system can correct this inefficiency, at least in part.

The development in the contributions within the economic analysis of law literature shows an increasing attention to the relationship between ex ante and ex post policies characterized by imperfections in their implementation, as complements or substitutes, both in providing the incentives to the optimal level of preventive care. So the debate about policy choice mainly focuses on the achievement of a given target in terms of efficiency in a framework where imperfections are considered as reasons to prefer one policy to the other.

Market-Based Policies: Taxes and Tradable Permits

Within the category of market-based instruments, there are policies that encourage behavior through market signals rather than through explicit directives to firms. So, in practice, rather than imposing uniform emission standards, market-based instruments introduce a cost for the firm in the form of a tax or of the price for a permit, leaving then the firms to deal with the problem to control and limit the pollution level on the base of their marginal costs.

So stressing an efficiency kind of argument, in the case of the implementation of market-based instruments, each firm determines until which level it is more convenient to reduce pollution given the possibility to pay a tax or to buy a polluting permit.

The two most important features of market-based instruments with regard to traditional command-and-control approach are cost-effectiveness and dynamic incentives for technology innovation and diffusion. On one hand, command-and-control policies, to set standards, require that policy makers obtain detailed information about the compliance costs, each firm faces the problem that such information may be not available to government; by contrast, market-based instruments provide for a cost-effective allocation of the environment control burden without the need for this information.

As market-based policies, the most common are tradable permits system and environmental taxes.

A system of tradable permits is based on the allocation of a number of permits to the firms, each of them allowing the emission of a given amount of a pollutant; if the facility is able to

reduce its emissions (preferably through the use of less polluting technologies), it can sell its remaining emission permits to other firms that are unable to meet their quotas.

It is clear that the advantage of this policy is the possibility to fix the level of pollution control and in the case of technological change, without additional government intervention, to freeze this level.

On the other hand, a policy based on taxation attributes a price on polluting activities that will be incorporated by the firm in the price of the products. In this case the incentive for the adoption of abatement techniques relies on the market mechanism because if a firm does not apply the optimal techniques, it will pay more taxes and sell its products at a higher price than other firms with negative effects in terms of competition.

Market-based instruments, as regulatory devices that shape behavior through market signals rather than explicit instructions on pollution control levels or methods, are often described as "harnessing market forces" because they can encourage firms and individuals to undertake actions that serve both their own financial interest and public policy goals (Stavins 1998).

Environmental taxes and tradable permits system are both market-based policy instruments, but their implementation is different: taxes fix the marginal cost for carbon emissions and allow quantities emitted to adjust, whereas tradable permits fix the total amount of carbon emissions and allow price level to change according to market forces. Because of these differences, the former are defined as "price" instruments for the correlated effect to increase the price of certain goods and services, thereby decreasing the quantity demanded, while tradable permits are defined as "quantity" instruments for the feature to directly fix the quantity trough the number of permits.

The literature on environmental policy choice describes alternative instrument taxes as price control instruments and tradable permits as quantity control ones, and many contributions compare their relative performance in terms of efficiency under uncertainty.

The starting seminal article of Weitzman (1974) analyzed the optimal instrument choice under a static partial equilibrium framework, consisting of a reduced form specification of costs and benefits from abatement. In the setup, an agency issues either a single price order (fixed price) or a single quantity order (fixed quantity), and these fixed policies result in different expected social welfare outcomes under uncertainty. Specifically, Weitzman shows that, with imperfect information about the abatement costs, the relative slopes of the marginal benefit (damage) function and the marginal cost function determine whether one instrument is preferred to another. If the expected marginal benefit function from reducing emissions is flatter than the marginal cost of abatement, then a price control is preferred. If, however, the marginal benefit function is steeper, then a quantity control is preferred.

In the law and economic literature, Kaplow and Shavell (2002) deal with the standard context of a single firm producing externality; moreover, they consider the case of nonlinear corrective tax, and multiple firms jointly create an externality, demonstrating the superiority of taxes to permits.

Despite the results of the majority of contributions that a taxation system is preferable to tradable permits system in terms of economic efficiency, this policy obviously faces political opposition. On the supply side of the market, companies oppose taxes, as a cost that implies a revenue transfer to the government and also as a factor that can imply negative effect on competition in an international context; on the demand side, consumers are typically not happy and pay at the end a higher price on the products, and environmental groups oppose taxes because, unlike tradable permits system, these fail to guarantee a particular reduction in the emission level.

Conclusive Remarks About the Mix of Policy Instruments

We have analyzed the different policies as alternative instruments that can be implemented to reach given environmental objectives considering, on a law and economic perspective, their different degree of efficiency.

But environmental policy instruments usually operate as part of a "mix" of instruments, and in

practice several different policies are applied to address a given environmental problem as broadly as possible with the target to cover all sources of pollution in every relevant sector of the economy.

The efficiency of these mixes can be enhanced by adhering to many of the same principles that guide the use of individual instruments and by explicitly considering the way in which different instruments interact.

For example, one possible mix could be tradable permits together with tax system.

On an efficiency point of view, while a policy based on "quantity," such as a tradable permits system, can provide a degree of certainty as to the environmental outcome, the compliance costs that will eventually be faced by polluters are likely to be quite uncertain under these systems. But this uncertainty can be reduced by introducing a tax system as a "safety valve" in the permit price. In effect, this allows polluters to emit whatever amount they like, in return for paying a fixed price, the "tax," for any emissions for which they do not hold an allowance, should the permit price exceed a predefined level.

This mix between environmental policies presents some economic advantages that are key motivating forces to try to develop research on this topic.

First of all, policy mix allows for exchanges across different systems and thereby facilitates cost-effectiveness, that is, achievement of the lowest-cost emission reductions across the set of linked systems, minimizing the overall cost of meeting the collective cap.

Mixed systems may also provide regulatory stability as an advantage for affected firms, in the sense that it may be more difficult to introduce changes in an emission-reduction scheme when those changes require some sort of coordination with other policies (Johnstone 2003).

There are also administrative benefits from the mix that come from sharing knowledge about the design and operation of different policies to find the best practice, but also from the reduction of administrative costs through the sharing of such costs and the avoidance of duplicative services.

Despite the just mentioned economic advantages, we cannot find in the law and economic literature until now so many researches that develop theoretical models based on the mixed use of different environmental policy instruments that are still considered mainly as alternative.

References

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

Cooter RD, Ulen TS (1988) Law and economics. Collins, Harper

Cropper ML, Oates WE (1992) Environmental economics: a survey. J Econ Lit 30:675–740

Epstein RA (1973) A theory of strict liability. J Leg Stud 2:151–204

Faure M, Skogh G (2003) The economic analysis of environmental policy and law – an introduction. Edward Elgar, Cheltenham

Johnstone N (2003) The use of tradable permits in combination with other environmental policy instruments. Report ENV/EPOC/WPNEP(2002)28/Final, OECD, Paris

Kaplow L, Shavell S (2002) On the superiority of corrective taxes to quantity. Am Law Econ Rev 4:1–17

Kolstad CD, Ulen TS, Johnson GV (1990) Ex post liability for harm vs. ex ante safety regulation: substitutes or complements. Am Econ Rev 80:888–901

Porrini D (2005) Environmental policies choice as an issue of informational efficiency. In: Backhaus JG (ed) The Elgar companion to law and economics, 2nd edn. Edwar Elgar, Cheltenham, pp 350–363

Rose-Ackerman S (1991) Regulation and the law of torts. Am Econ RevPap Proc 81: 54–58

Shavell S (1984) Liability for harms versus regulation of safety. J Leg Stud 13:357–374

Shavell S (1986) The judgement proof problem. Int Rev Law Econ 6(June):45–58

Stavins R (1998) Market-based environmental policies. Resources for the Future, Discussion Paper 98–26

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

Equilibrium Theory

Pasquale L. Scandizzo
CEIS (Centre for International Economic Studies), University of Rome "Tor Vergata", Rome, Italy

Abstract

Equilibrium is a key concept of modern science, from classical mechanics to biology, so

that its importance for economics should not be a surprise. More surprising is perhaps its central role in the tentative transition of economics from an essentially institutional set of disciplines to a unified branch of modern science, based on a rigorously axiomatic system. The main attempt for such a transition, which has as protagonists several Nobel prizes, was consumed in the 1950s, in an atmosphere first of elation and then of disillusion that appears very similar to that experienced 20 years earlier by mathematicians because of the failure of the Hilbert unification program. In spite of the somewhat spectacular nature of both successes and failures of its mathematical theory, however, general equilibrium has a central role in the history of modern economics that goes much beyond its formal treatment. This role is inextricably related to the notion of the market and is perhaps the primary constituents of the eclectic nature and vitality of contemporary economics.

General Equilibrium as a Property of the Market

Perhaps because of its increasing importance to the modern economy, the concept of market equilibrium has been evolving over the course of the story, so as to embrace different categories, in some cases more inclusive and in others more specific, which extend over the entire arc of economic phenomena. This extension of a concept originally defined in a very simple fashion has led to considerable confusion, especially since the theory and practice of modern economic systems are, in fact, completely dependent on market equilibrium as a category of the spirit, even more than as a theoretical model.

What then is a market equilibrium and in what sense can it be defined as a "general equilibrium"? The word market was born as the past participle of the Latin verb mercari, which means trading, and designates the beginning of its use of a physical place where the goods available for exchange were exposed and where, therefore, negotiations and exchanges were carried out. The designation of the place, still widely used in everyday language, came to denote at the same time the space, the exchange, and the agents engaged in trade of specific goods. As a result, the place evoked by the term became increasingly virtual, the transition between the material and the virtual occurring through the discovery of the properties of the markets that extend well beyond their apparent spatial limits and culminate in the concept of market equilibrium. Equilibrium, on the other hand, was a concept that only gradually came to characterize the market, as its assertion went hand in hand with the increasing virtuality of the exchange and the agents involved in making it happen.

The process by which the market became dematerialized, in fact, can be seen as an increasing perception of its role as a mechanism to balance the exchange, by ensuring that agents meet and act appropriately to achieve a balance of actions and desires. For example, one of the precursors of economic liberalism Richard Cantillon (1755) offers a description of the market and its reasons to demonstrate how the activity originating from the establishment of a periodical fair creates broader consequences:

> There are some villages where markets have been established by the interest of some proprietor or gentleman at court. These markets, held once or twice a week, encourage several little undertakers and merchants to set themselves up there. They buy in the market the products brought from the surrounding villages in order to carry them to the large towns for sale. In the large towns they exchange them for iron, salt, sugar and other merchandise which they sell on market days to the villagers. Many small artisans also, like locksmiths, cabinet makers and others, settle down for the service of the villagers who have none in their villages, and at length these villages become market towns. A market town being placed in the centre of the villages, and at length these villages become market towns. A market town being placed in the centre of the villages whose people come to market, it is more natural and easy that the villagers should bring their products thither for sale on market days and buy the articles they need, than that the merchants and factors should transport them to the villages in exchange for their products. (1) For the merchants to go round the villages would unnecessarily increase the cost of carriage. (2) The merchants would perhaps be obliged to go to several villages

> before finding the quality and quantity of produce which they wished to buy. (3) The villagers would generally be in their fields when the merchants arrived and not knowing what produce these needed would have nothing prepared and fit for sale. (4) It would be almost impossible to fix the price of the produce and the merchandise in the villages, between the merchants and the villagers. In one village the merchant would refuse the price asked for produce, hoping to find it cheaper in another village, and the villager would refuse the price offered for his merchandise in the hope that another merchant would come along and take it on better terms. (Cantillon 1755)

For Adam Smith, one of the founders of the notion of equilibrium, the market was also originally identified in a physical space, such as a place dedicated to trade in the city or even the entire city (as a market town) with respect to the surrounding countryside. For him, however, the essential feature of the market was its ability to feed and be fed by the division of labor. In this sense, the market becomes a place dedicated to the determination of value by a threefold change: (a) from primitive societies to modern societies, (b) from the family to the firm, and (c) from the countryside to the cities. Smith's conception of the market is therefore first and foremost ideological, as pointed out by Joan Robinson (1962). It identifies the transition between two ideal states "the economy of barter" and "the market economy," as a form of substantial progress that is the basis of value creation and prosperity of modern societies. In this sense, Smith sees the market as an anthropological construct, which explains the redemption of man from the "rough state" where the assets and exchanges are both limited by the absence of the division of labor.

Ricardo, who starts from the paradigm of the Smithian division of labor to develop his fundamental theory of comparative advantage, seems to show a total nonchalance towards the concept of the market as compared to that of equilibrium. The market as such is evoked only from time to time, as a place dedicated to trade, in which the goods produced are carried. However, in the great debate on international trade, it remains a kind of constituency space, which distinguishes the "foreign" from the "home" market, both defined almost as a side effect of trade between countries. In fact, both Smith and Ricardo did not seem to attach much importance to the market, in the sense that they take its presence for granted within a system that they characterize more than as a "market economy," as an economy of capital. Yet the fundamental problem, in this system, is not to explain the nature and behavior of economic entities, but rather to solve the problem of achieving equilibrium through the formation of value and the distribution of wealth.

> [. . .] On the contrary, as the rise in the real value of silver, in consequence of lowering the money price of corn, lowers somewhat the money price of all other commodities, it gives the industry of the country where it takes place, some advantage in all foreign markets, and thereby tends to encourage and increase that industry. But the extent of the home market for corn must be in proportion to the general industry of the country where it grows, or to the number of those who produce something else, to give in exchange for corn. But in every country the home market, as it is the nearest and most convenient, so is it likewise the greatest and most important market for corn. That rise in the real value of silver, therefore, which is the effect of lowering the average money price of corn, tends to enlarge the greatest and most important market for corn, and thereby to encourage, instead of discouraging, its growth. (Ricardo 1821, Chapter 24, p. 42)

Walras was to argue, with the esprit de finesse of his national tradition, that the market was not a single place, but a system of interdependent markets, and therefore, interdependence was its constitutive principle which also linked the participants. From the subjective point of view, the market consisted then of entrepreneurs, brokers, and owners of factors of production (workers, capitalists, and rentiers) connected with each other and with the goods exchanged by a mechanism (the tâtonnement) search of equilibrium prices. Walras does propose a unique identification of entrepreneurs, not so much because, according to his theory, they maximize their utility through the pursuit of maximum profit, but because they are characterized as intermediaries between the markets for factors of production and that of the goods. Under equilibrium conditions of production, the entrepreneur does not get either profits or losses (Walras 1877b, p. 232, 1954, p. 225). But equilibrium, for Walras, is actually

a theoretical notion that characterizes the market: it constitutes the normal state to which all variables tend perpetually and automatically in a competitive free economy (Walras 1954, p. 224). Since it contains implicitly the equilibrium of exchange, the market owns the further property of equality between the supply and demand for final and intermediate goods, as well as factors of production.

After Walras, it was Marshall who addressed more explicitly the problem of equilibrium, by characterizing the market as a meeting place of supply and demand. Although based on the comparison of "needs" or "desires" (wants), rather than the incipient neoclassical paradigms (demand, supply, perfect competition), this concept was somewhat overshadowed by the theories of trade in Smith and Ricardo. It needed, however, to be made explicit and descriptive, to accommodate the concept of price formation, bargaining power, and competition. In Book V of his *Principles of Economics*, Marshall begins by tracing a continuum of markets, wider or narrower depending on the position between the two extremes of easily transferable products from open markets and relatively immobile products captured by closed markets:

> Thus at the one extreme are world markets in which competition acts directly from all parts of the globe; and at the other those secluded markets in which all direct competition from afar is shut out, though indirect and transmitted competition may make itself felt even in these; and about midway between these extremes lie the great majority of the markets which the economist and the business man have to study. (Marshall, *Principia*, Libro V, Cap. I)

In this continuum, the unique characteristic of markets is that they allow economic agents to find a balance between desires and efforts. The easiest way to find this balance, says Marshall, tracing in this way the classical phylogeny, is to a person who gets what she wants directly from her work. A second tool is the barter, but both of these mean being primitive and limited in their practical consequences; in the end, the emergence of the market appears to be a sheer necessity.

In the market two quantities are compared, subject to variation: the supply price, given by the minimum price that the agent who proposes the sale is willing to accept for the goods offered, and the bid price, which is the maximum price that the agent who contemplates the purchase is willing to pay. Generally, if the goods are divisible, there are no particular problems in finding the equilibrium point, i.e., the quantity of goods exchanged for which the bid price and the supply price match. There are, however, particular markets where the achievement of this point is not trivial: the labor market, for example, is characterized by the fact that for anyone who offers the sale, only one unit has to be put on the market and may be in need, so that the worker may be willing to accept a very low salary. Her supply price, determined on the basis of the cost of her effort, may not coincide with the price for which she is forced to sell her work to survive.

Marshall's treatment is striking for its clarity but also, as in the case of the classics, for the essentially passive nature attributed to the market. The latter, in addition to be configured as the virtual space of the exchanges, i.e., as a locus operandi with desirable characters of full information, competitiveness, and so on, does not appear to have in any way a coordinating role. The coordination of trade is in fact an automatic consequence of the comparison between supply and demand (or, through the supply price and the bid price, the comparison between effort and desire) and does not even need to resort to the Walrasian auctioneer. In fact, the very concepts of demand and supply prices appear as abstractions, raising the question of balance in an essential way, and can be considered the disembodied protagonists that allow to bypass a theory of the functioning of the market. This is perhaps also due to the fact that, as noted by Colander (1995), Marshall was aware of a higher complexity of the general equilibrium problem and that his conception was more subtle and pervasive than Walras, combining the ideas of multimarket coordination and equilibrium "generality" with those of dynamic adjustment and coordination from a plurality of institutions. While the Walras model is addressed to the solution of a multiple matching problem between goods supplied and goods demanded in an interdependent perspective, Marshall

conceives general equilibrium as a process with "chaotic" characteristics. This process arises from the fact that preferences are constantly updated and economic agents strive to find the equilibrium again and again by moving along "corridors" of coordination which are contingent on the existence and operability of institutions.

A little less than a century later, in a text characterized by the almost total absence of the word "market," Thorstein Veblen constructs a theory of the entrepreneur, based on the idea that general equilibrium is a dynamic process, depending on the figure of the "captain of industry," motivated by the opportunities provided by the market to acquire a monopoly position in a economy dominated by machines and processes of product standardization.

Veblen's model of the "captain of industry," dear to the frontier capitalism, has the reference specimen of big businessmen, such as the railway investors, whose fortunes, however, consisted not only in the vision and passion for the gains but also in the ability to fight with success against other entrepreneurs and, on the side of consumers, seize the best opportunities for emerging markets. To indicate the wisdom of this opportunistic attitude, Veblen cites, in particular, the saying "charging what the traffic can bear" (Veblen 1904, Chapter 3), which refers to the nascent industry of American Railroads – a service rather than an industry – but characterized by the use of powerful machines.

The market and with it the very notion of equilibrium thus tend to disappear in a heroic perspective of entrepreneurship. Veblen anticipates what will be the central observation of Coase and the starting point of the neo-institutional economics, which identifies the company as an alternative to the market and its command and control structure as an alternative to the impersonal balance between demand and supply. Physiocrats, classical and neoclassical economists, and, in their specificity, even the members of the Austrian school, from Mengel to Böhm-Bawerk, in fact, consider the entrepreneur and the enterprise an integral part of the market, as opposed to the family economy and the model of consumption. In Veblen, however, the idea of a self-governing equilibrium reappears in the role of the captains of industry as of those who, fighting with other entrepreneurs, to eliminate them from the market, achieve equilibrium by freeing the economy from an excess of management, because "[. . .] it appears that the greater the amount of pecuniary management, the smaller are the services provided to the community."

The heroic function of the entrepreneur for Veblen coincides with a form of competition that seeks the disequilibrium and is therefore completely different from that of Walras and Marshall. The experience of American capitalism will inspire 20 years later Max Weber, in attributing a role equally dynamic and heroically ascetic to the moral entrepreneur created by the Protestant Reformation. It is interesting to contrast the points of view of these two authors, Veblen, an economist with sociological propensities, and Weber, a sociologist with inclinations as an economist, because they derive a heroic role for the entrepreneur from completely different views of his moral and creative features. For Veblen, the entrepreneur is spurred by the profit motive and the desire to amass a fortune. The drive towards these goals has its social utility, but to pursue them effectively, the entrepreneur must be substantially amoral and essentially disruptive. For Weber, on the contrary, the success of the capitalist entrepreneur embeds, through the Protestant ethic, moral character and a social quality dramatically higher for the market before birth and consolidation of the reformed religion.

More recently, Kirzner, the most influential living economist of the Austrian school, takes up the theme of the Socratic role of the entrepreneur, making him, however, the principal agent of the search for market equilibrium. According to Kirzner (1973), the function of the entrepreneur as an innovator gives impetus, through her tireless spirit of initiative, to the entire economy. In doing so, the entrepreneur plays a crucial role in correcting market failures, through the activities of arbitrage and speculation. In the process of finding and reaching, the entrepreneur achieves general equilibrium in a far broader sense, as balance, dynamics, and completeness of the markets. In addition, speculative activities, often held to blame, cause substantial benefits for

consumers, because they eliminate the rents due to disequilibrium and push the market towards more efficient conditions. For the entrepreneur to be able to fully release her energy in a beneficial way, it is necessary that the institutional environment is suitable, and entrepreneurial initiative is not mortified by the "traps and snares" of bureaucracy and regulation.

The entrepreneur is therefore the keystone of the economic system because she keeps alive the work, as in the heroic vision of Veblen and Weber, and at the same time, through her constant pursuit of profit, ensures that market activity goes to fruition. As an innovator, Kirzner (1989) argues that the entrepreneur also operates ethically even when she captures high incomes and excess profits. These are in fact the result of the fact that she has discovered the use of resources, thus making to come to life (economic) economic goods that did not exist before. As a beneficiary of distributive justice that rewards merit, the entrepreneur is therefore a fully consistent moral subject.

In the face of these attempts to explain the historical and institutional market, the long wave of the neoclassical school, which is rooted in the principle of specialization and cooperation, reemerges with great force. Wicksteed is perhaps the economist who argues more persuasively, without the use of mathematics, the paradigm and the marginalist theory of equilibrium which will become the core of the so-called neoclassical synthesis. His Presidential Address (1914) contains a concise summary of the marginalist theory of value and distribution and, at the same time, offers an incisive way in the implications of this theory for the nature of the institutions. First of all, the economy is based on a principle of cooperation: "The economic body [. . .] of an industrial society is an instrumentality for which each man, doing what he can for some of his peers, he gets what he wants from others". Secondly, the principle of maximization does not imply global complex calculations, but only a comparison algorithm of local differences:

> [. . .] which means, in common parlance, that what an individual will be willing to give something in return, but be rather lacking, is determined by comparison of the difference that he thinks that its possession would do for him except that he would give anything in return would cause [. . .].

Moreover, it is this principle of equality of differences that explains both the balance of the individual and that of the market, i.e., between individuals, in the sense that as long as someone believes that an act of exchange may reduce the gap between two opposing differences, he will attempt to make the exchange [. . .].

E

What Is General Equilibrium?

Because of its intrinsic complexity and its very diverse history and in spite of its centrality in economic theory, the notion of general equilibrium that has gradually achieved consensus is both subtle and controversial and still subject to alternative interpretations by different scholars. The point of departure concerns the more general notion of "equilibrium," where, since the physiocrats, the main governing idea is the objective property of a material "balance" between demand and supply of both goods and services. It is this balance, which is surprisingly attained by exchange through unknown and somewhat mysterious means (e.g., Smith's "invisible hand"), that ensures that the economy is not clogged with unwanted stocks of commodities or plagued by involuntary unemployment or lack of goods or services needed or desired. Material balance generates the further idea that a second, essential property of equilibrium has a subjective nature: in equilibrium agents are content in the sense they feel themselves in a desirable state of balance, because they sell what they want to sell and buy what they want to buy at mutually agreeable rates of exchange. This subjective condition in turn leads to stipulate that in equilibrium agents realize their plans and are satisfied because they behave according to what they believe is right to pursue their objectives and have no need or desire to change.

So far the notion of equilibrium described does not have any special connotation as partial or general, but simply characterizes a condition where things are at rest, in the sense that there is no incentive to change the position of any agent,

either to reduce objective costs or to increase subjective satisfaction, unless underlying exogenous conditions change. In this context, the "partial" equilibrium model emerges from a desire to simplify the analysis of price formation, limiting the interaction between demand and supply and the role of prices to one market or to a subset of interdependent markets. Partial equilibrium, therefore, does not mean that the multiplicity of markets is necessarily ignored nor that any of the properties of equilibrium (material balance, subjective self-fulfilling plans) are discarded but only that one part of the economy is considered exogenous. More specifically, partial equilibrium generally ignores the linkage between price determination and income formation, as well as the wealth effect caused by the fact that the increase or decrease in prices changes the value of agents' endowment of economic resources.

A further, often tacit, characterization of market equilibrium is that it is "competitive," i.e., it arises in a market where there is a plurality of agents, each too small to determine, through her individual behavior, the outcomes of the exchange. Of course, monopolistic equilibrium has been cultivated as a theoretical field by several scholars, but it is typically the solution of a noncooperative game and does not possess the combination of objective (material balance) and subjective characteristics of competitive equilibria. As John Nash (1950, 1951) demonstrated, in fact, the notion of subjective equilibrium does not go hand in hand with the accomplishment of material balance. In the theory of noncooperative games, a Nash equilibrium is reached if each player maximizes her objective function under the full knowledge of the strategies of the other players. In such an equilibrium, a higher level of mutual rest is reached in that no player may benefit by changing her strategy if no other player does so. This does not mean, however, that players would not like to change the allocation of goods or services characterizing the equilibrium. In most cases they would, but they cannot, because such an allocation is determined by a -higher-level equilibrium among the strategies leading to the allocation. A Nash equilibrium is thus a meta-equilibrium, and while a competitive market equilibrium qualifies as a Nash equilibrium, a Nash equilibrium is not necessarily a competitive equilibrium and does not necessarily require material balances between demand and supply or agents' contentment about the allocations realized by the enactment of their strategies.

What is then "general equilibrium"? One is tempted to reply that general equilibrium describes a condition where all markets are in equilibrium, both in the sense that all markets are cleared (demand equals supply) and all agents fulfill their plans. However, while this definition certainly appears simple and direct, it is not complete, since general equilibrium, unlike partial equilibrium, in addition to material balances and subjective fulfillment requires that the distribution of wealth is consistent with resource allocation. General equilibrium theory, in fact, concerns three different circles of causation: (i) between demand and supply of goods and services on one hand and prices and incomes on the other, (ii) between the formation of incomes from demand and supply of factors of production and their prices, and (iii) between the initial resource endowment and the redistribution caused by productive choices and institutional transfers. The precise way in which these three circles interact was not clear until in relatively late times, R. Stone and A. Brown (1962) formalized it in the so-called social accounting matrix (SAM). The focus on the proof of the existence by the celebrated Nobel Prize-winning economists Kenneth Arrow and Gerald Debreu, furthermore, while in itself a benchmark of economic theory, for a long time diverted the attention of the scientific community from more practical features of general equilibrium, such as its computability under alternative scenarios of competition and information and its use for planning and project evaluation. Computable general equilibrium (CGE) models, mainly developed as a result of research efforts at the World Bank in the 1970s, were thus mainly a spin-off of the application of SAM and even in their advanced, present-day form, they tend to evoke the initial dualism between a core set of social accounts and a complementary and highly variable set of behavioral and technical equations.

While at first it may not even be seen as the same subject, since the beginning of its being theorized, the outcome of general equilibrium has been identified with efficiency, in the sense that a fully competitive price system brings about more efficient outcomes than any other practical arrangement or planning mechanism. The linkage between competitive equilibrium and efficiency was formalized by Pareto (1909) and Bergson (1938) and, in its most advanced form, by Arrow (1951) and Debreu (1951) in the form of a full equivalence theorem. Pareto's definition – exemplar in its efficacy – characterizes efficiency as a feasible allocation, i.e., the choice of feasible sets of goods to consume and to produce on the part of the economic agents, that cannot be changed without making at least one agent worse off. Efficiency thus simply implies that it is not possible to modify an efficient allocation without damaging someone. It is the expression of the social trade-off that arises once all possible improvements with no impact on distribution have been made.

Two basic theorems of welfare economics link Pareto efficiency to the notion of general (Walrasian) equilibrium. The first says simply that any general equilibrium is Pareto efficient (Arrow 1951; Debreu 1951), while the second theorem states that under certain, rather general, conditions, given a Pareto-efficient allocation, it is possible to find a corresponding general equilibrium and, in particular, a supporting price vector. These two theorems have important policy implications, because they suggest that Pareto effecirncy may be obtained either through a decentralized market system (as in the Walrasian model), by letting prices be freely determined by the exchange mechanism, or as a planned allocation, by imposing a system of prices that will support it.

The relationship between Pareto efficiency and general equilibrium foreshadows the question of its existence, which was taken up by Arrow and Debreu (1954), as well as McKenzie (1959) in a series of celebrated contributions. Their work, although focused on a rather narrow subproblem, essentially proved that under certain conditions aggregate demand and supply could be equated by a set of nonnegative prices. This was a nontrivial result that was contingent on a series of rather restrictive assumptions but was obtained through a mathematical powerful and unifying instrument (the fixed-point theorem) that was in itself shining for originality and simplicity. The result had two drawbacks, however. First, it did not cover nor it proved to be a feasible base for finding circumstances under which the equilibrium was unique. Second, as proved in a series of important and somewhat astounding later contributions by Sonnenschein (1972, 1973), Debreu (1974) himself, and Mantel (1974), the base of the existence proof was a construct, named "the aggregate excess demand function," which, although resulting from the aggregation of individual demand and supply, was not bound by the limitations deriving from the postulates of rationality. In what has been called "the everything goes" conclusion, in fact, it was proved that such a function, even though the result of individual rational behavior, is not characterized by any special mathematical property. Thus, the existence of general equilibrium seemed to be quite independent of its "micro-foundations," as a consequence of an essential weakness of the microeconomic "rationality" assumptions, which were proved to be not sufficiently discriminating to impose anything resembling rationality on aggregate behavior.

General equilibrium theory, however, was not devoid of consequences for applied economics. In a series of important research attempts, in large part conducted at the World Bank, several generations of computable general equilibrium (CGE) models were developed and gradually became an important and useful tool for policy analysis. In these models, social accounting matrices (SAM) became the core of the representation of general equilibrium as a circular flow of production, consumption, and incomes, with prices in all markets as the equilibrating variables. Solving algorithm started with fixed-point (Scarf and Hansen 1973) and mathematical programming procedures (Norton and Scandizzo 1981; Walbroeck and Ginsburg, 1981) and gradually developed into nonlinear equation systems and local or global search solution methods (Devarajan et al. 1997). At present, while the macroeconometric models

prevailing in the 1970s have all but disappeared from the economic practice, CGE models are increasingly used around the world, in both their static and dynamic versions, as tools to analyze economic policy options.

Conclusions

General equilibrium has accompanied the development of economic theory both as a founding concept and as a scientific challenge. Its classical version interprets the market as being characterized by a search for coherence, fulfillment, and balance, from the point of view of both the agents and the goods and services traded. In this interpretation, the historical notion of the market as a place in space and time is progressively changed into that of a virtual space, whose prevalent feature is to harbor the dynamism rather than the locus of exchange. The neoclassical view continues this process of dematerialization of the market and pushes it further by focusing on the abstract nature of equilibrium as a mathematical construct and on its existence and uniqueness properties. The ambition of this second approach, which permeates the founding phase of mathematical economics, is to unify economic theory by providing a general and rigorous connection between microeconomic and aggregate behavior.

While successful in imposing a method of theoretical investigation that remains one of the assets of modern economic theory, the neoclassical program did not go beyond proving a theorem of existence of general equilibrium and of seemingly important but somewhat rarified properties of socially efficient allocations. Embarrassingly, the program also proved the micro-foundation illusory, as a consequence of an intrinsic incapacity of the rationality assumptions to characterize aggregate behavior. Surprisingly, general equilibrium theorizing has been instead more successful in applied economics, by providing, through the social accounting matrix (SAM) and computable general equilibrium (CGE) models, an apparatus of sufficient rigor and sophistication to systematize data collection and frame the analysis of complex economic policies.

References

Arrow KJ (1951) An extension of the basic theorems of classical welfare economics. In: Neyman J (ed.), Proceedings of the second Berkeley symposium on mathematical statistics and probability. University of California Press, Berkeley

Arrow KJ, Debreu G (1954) Existence of an equilibrium for a competitive economy. Econometrica 22(3): 265–290. https://doi.org/10.2307/1907353. https://doi.org/10.2307%2 F1907353

Bergson A (1938) A reformulation of certain aspects of welfare economics. Q J Econ 310–334

Cantillon R (1755) Essai sur La Nature de Commerce en Géneral (ed: Hibbs H). Macmillan, London, 1931

Colander D (1995) Marshallian general equilibrium analysis. East Econ J 21(3)

Debreu G (1951) The coefficient of resource utilization. Econometrica 19(3):273–292

Debreu G (1974) Excess-demand functions. J Math Econ 1:15–21. https://doi.org/10.1016/0304-4068(74)90032-9. https://doi.org/10.1016%2 F0304-4068%2874%2990032-9

Devarajan S, Go D, Lewis J, Robinson S, Sinko P (1997) Simple general equilibrium modeling. In: Francois J, Reinert K (eds) Applied methods for trade policy analysis. Cambridge University Press, Cambridge, UK

Kirzner IM (1973) Discovery capitalism and distributive justice (Oxford 1989)

Mantel R (1974) On the characterization of aggregate excess-demand. J Econ Theor 7:348–353. https://doi.org/10.1016/0022-0531(74)90100-8. https://doi.org/10.1016%2 F0022-0531%2874%2990100-8

McKenzie LW (1959) On the existence of general equilibrium for a competitive economy. Econometrica 27(1):54–71. JSTOR 1907777

Nash JF Jr (1950) Equilibrium points in n-person games. Proc Natl Acad Sci U S A 36:48–49

Nash JF Jr (1951) Non-cooperative games. Ann Math 54:286–295

Norton R, Scandizzo PL (1981) General equilibrium computations in activity analysis models. Oper Res 29(2):243–262

Pareto V (1909) Manual of political economy (Trans: Kelley AM). New York 1971

Ricardo D (1821) On the principles of political economy an taxation, 3rd edn (ed 1821). John Murray, London

Robinson J (1962) Economic philosophy. Transaction Publisher

Scarf HE, Hansen T (1973) The computation of economic equilibria, vol 24, Cowles Foundation for Research in economics at Yale University, Monograph. Yale University Press, New Haven/London

Sonnenschein H (1972) Market excess-demand functions. Econometrica (The Econometric Society) 40(3): 549–563. https://doi.org/10.2307/1913184. https://doi.org/10.2307%2 F1913184. JSTOR 1913184

Sonnenschein H (1973) Do Walras' identity and continuity characterize the class of community excess-demand

functions? J Econ Theor 6:345–354. https://doi.org/10.1016/0022-0531(73)90066-5. https://doi.org/10.1016%2 F0022-0531%2873%2990066-5
Stone R, Brown A (1962) A computable model for economic growth. Cambridge Growth Project, Cambridge, UK
Veblen T (1904) The theory of business enterprise. Scribner New York
Walbroeck JL, Ginsburg VA (1981) Activity analysis and general equilibrium modelling. North Holland Publishing, Amsterdam, Webley, O. 1986
Walras L (1954) Éléments d'économic politique pure, elements of pure economics (trans: William Jaffé from the 1926 ed). Richard D. Irwin, Homewood
Warlas L (1877b) Éléments d'économic politique pure, 2nd part Ist ed. L.Cobraz, Lausanne
Wicksteed PH (1914) The scope and method of political economy in the light of the "marginal" theory of value and of distribution. Econ J 1–23

Essential Facilities Doctrine

Frédéric Marty
CNRS - UMR 7321 Gredeg, Université Côté d'Azur, Université Nice Sophia Antipolis, Valbonne, France

Definition

The essential facility doctrine is a disputed concept in the field of competition law enforcement. According to this doctrine while a dominant operator controls an asset that its competitors cannot bypass to access the market because of its natural monopoly situation or because the unreasonableness of its replication in financial or in technical terms, it may be bound to provide them an access in fair, reasonable, and non-discriminatory terms. This approach may lead to far reaching remedies and is all the more challenged that it is also implemented to intangible assets.

Introduction

The essential facilities doctrine (hereafter the EFD) stems from the notion of market failure and, even more specifically, from the concept of natural monopoly. The EFD may be used in a situation in which an economic operator access to the market exclusively depends on the decision of a facility owner without any available alternative. The facility owner may be one of its competitors in the downstream market or a market player who controls an infrastructure which plays as a bottleneck to enter the market. The facility at stake can be an infrastructure, an upstream product or service, or an intangible asset. In other words, the facility owner can be a network operator, an upstream monopolist, or a patent holder. The EFD may be activated within the scope of competition law if this upstream monopolist's refusal to grant an access can be qualified as an exclusionary or an exploitative abuse. It can also be used within the scope of antitrust law, at least in principle, if this refusal participates to a strategy aiming at monopolizing a downstream market.

The denial of access may be an absolute one – it is a clear cut refusal to deal – or a "relative" one. In such a case, the access is still possible but the facility owner imposes unfair or discriminatory conditions in terms of access charges or provides to its downstream competitors an access characterized by deteriorated technical conditions that impair the quality of the service provided to the final user. Such an altered access distorts competition.

The first situation may be commonly observed in deregulated industries. A former State-owned vertically integrated firm, controlling a natural monopoly segment (as the local loop in a telecommunication network, the transportation infrastructure in gas or electricity sector. . .) may impose excessive access charges to its new competitors in the downstream market. Margin squeeze strategies may be at stake in such situations. They may consist in imposing an excessive price for the upstream service and in charging an exclusionary price for the downstream one. Cases as *Deutsche Telekom* (case 37.451, European Commission, 21 May 2003) illustrate this type of strategy, implemented in this instance to impair the access to the market of new competitors (ADSL Internet access providers). The incumbent leverages its dominant position to downstream

competitive markets thought cross subsidizations or by raising its rival costs. Imposing excessive access prices may also be a way to accumulate profit margins to finance exclusionary prices in an associated market in a diversification strategy.

The second situation, characterized by voluntary and artificially degraded technical conditions of access, may be embodied by the *Trinko v. Verizon* decision of the US Supreme Court (540 U.S. 398, 2004). In this case, the incumbent, who controls the essential facility, artificially deteriorated the quality of the service provided to final users by its downstream competitor (here AT&T). In other words, its competitor cannot guarantee to its users the same quality service than the one the incumbent provides to its own customers on the downstream market. The access was not denied but the competition on the merits was structurally distorted.

We first consider the implementation of the EFD to network infrastructures in the USA, before analyzing in a second part its use in the EU context, both for tangible and intangible assets. Our third part opens a discussion.

No Man Is a Prophet in His Own Country: The EFD Under Intense Criticisms from US Scholars

The EFD was initially implemented while a "natural monopoly" infrastructure is at stake. Firstly, we assess to what extent this doctrine may be activated in network industries to impose a mandatory access in fair, reasonable, and not discriminatory conditions to the benefit of the downstream competitors of a vertically integrated former monopoly. Secondly, we consider the distinction between the cases of competition laws and of sector-specific regulations. Its implementation as a regulatory tool in the United States contrasts with the situation in the field of antitrust laws. While the concept was crafted within the scope of the Sherman Act enforcement, its activation within the scope of antitrust laws is increasingly challenged.

Reluctances to Implement the Concept in the Legal Sphere

The EFD was crafted in the field of network industries in the USA in the early twentieth century. Its first use was related to the natural monopoly issue. As a competition through the infrastructures is technically impossible (or collectively suboptimal in terms of welfare), it is necessary to make possible a competition through the services. It supposes to guarantee a fair, undistorted, and equal access to the natural monopoly to any competitor wanting proposing its services to the final consumers, in the downstream market, in which the competition is possible. The most archetypal implementation of this approach was performed in the 1980s in the telecommunications sector with the AT&T decision.

Promoting a downstream services-based competition despite the existence of an upstream monopolistic bottleneck both implies network open access architecture and a regulation that ensures the effectiveness of a *level playing field*. The 1982 US Department of Justice consent decree in the AT&T case embodied such a logic (United States District Court of the District of Columbia, US v AT&T Co., 552 F. Supp. 131, August 1982). If this case led to the incumbent break up, the key stone of the remedies consisted in a mandatory and undistorted access to the network guaranteed for all the *baby bells,* e.g., for all the downstream competitors in the telecommunication services markets. In the case at hand, the EFD was implemented in the framework of a market building strategy in the context of the sector liberalization. Such a strategy is also at stake through the EU directives promoting a structural unbundling of vertically integrated State owned monopolies in field of utilities. However, the EFD was not crafted in the field of network industries deregulation but as a competition tool 70 years before.

Indeed, the first occurrence of this concept can be found in the US Supreme Court case law in *Terminal Railroad Association* (224 U.S. 383, 1912). The Supreme Court admitted that while the unification of railroad facilities in Saint Louis, Mississippi, was permissible, considering the efficiency gains it produces, it remains that a refusal to grant access to a competitor to these ones may be analyzed as an antitrust law infringement as it discriminatorily impairs its capacities to access to the market.

Thereupon the agreement among competitors can be analyzed as a combination in the purpose to restraint interstates trade. Throughout the twentieth century several US courts decisions implemented this concept to unilateral and coordinated practices. According to this case law, an access to a bottleneck may be compelled as soon as a competitor has no alternative to enter the market. It was for instance the case in *Associated Press* (326 U.S. 1, 1945), *Lorrain Journal* (42 U.S. 143, 1951), *Otter Tail* (410 U.S. 366, 1973), *Hecht Football* (570 F.2d 982, D.C. Cir., 1977), or in *Aspen Skiing* (427 U.S. 585,1985).

However, the concept of essential facility, and more precisely its implementation within the scope of Antitrust laws, was formally rejected in harsh terms by the US Supreme Court in 2004 in its *Trinko* ruling. Quite surprisingly, according to the Supreme Court, the essential facilities doctrine was never endorsed by its own case law. Insofar as it is not an antitrust concept, it may be used, at best, in the field of sector specific regulation. . . If the US Supreme Court overturned its own long-standing jurisprudence (from 1912 to 1985), it may be related both to the specific conditions of this case (a class-action brought by a law-firm searching for damages) and to the evolution of the academic literature regarding this issue since the 1960s. Indeed, legal scholars have increasingly seen in the implementation of EFD a violation of fundamental rights as property rights or freedom to contract (Areeda 1990).

Such a mistrust is not so surprising considering other courts decisions both from the *Lochner* era (the conservative US Supreme Court case law before the progressive turn of the late 1930s' and the post war *Warren* era) and from the *Chicagoan* turn of the late 1970s. The Supreme Court stated in *Colgate* (250 U.S. 300, 1919) that a company has the power to choose with whom to contract and to set its conditions. Since the market power does not result from a *monopolization* strategy, antitrust laws have not to impose to contract with any other undertaking. In 1980, the *Berkey Photo v. Eastman Kodak* decision (444 U.S. 1093, 1980), confirms that a monopoly, since its position results from its own the merits, can charge the price it decides.

These two judgments have to be placed into in their respective proper contexts. The first one is representative of the *classical legal thought.* Neither antitrust laws nor public regulations have to impair fundamental rights. The second one is typical from the *Chicagoan* approach of antitrust laws enforcement. The *GTE Sylvania* decision (Continental Television v. GTE Sylvania, 433 U.S. 36, 1977) widened the scope of the rule of reason at the detriment to per se prohibitions. The *Reiter Sonotone* decision (Reiter v. Sonotone Corp., 442 U.S. 330, 1979) endorsed the consumer welfare approach. The *Eastman Kodak* decision confirmed the *Chicagoan* view according to which a monopoly can lawfully extracts the competitive rents it had created. In this framework, Antitrust laws only have to sanction the extension of a market power on other basis than the merits. Extracting the surplus created is considered both as legitimate and as desirable in terms of market dynamic (Carlton and Heyer 2008).

Indeed, the Sherman Act, promulgated in 1890, does not prohibit the mere possession of a monopoly power. A monopoly position is not an antitrust issue in and of itself while its acquisition and its maintenance stem from the undertaking's own merits. As the Supreme Court stated in *Grinnell Corp* (384 US 563, 1966) "The offense of monopoly under § 2 of the Sherman Act has two elements: (1) the possession of monopoly power in the relevant market and (2) the willful acquisition or maintenance of that power as distinguished from growth or development as a consequence of a superior product, business acumen, or historic accident." As a consequence, the notion of merits goes far beyond the scope of considerations narrowly related to past investments or to the risks initially taken by the incumbent.

Such a view contrasts with the underlying logic of the Judge Learned Hand's ruling in *Alcoa* (148 F.2d 416, 2d Cir. 1945). According to the current dominant conception of antitrust laws enforcement, a monopolist is not bound to adjust its market behavior in order to guarantee a *living*

profit to its competitors. On the contrary, Learned Hand considered that a vertically integrated undertaking that enjoys a monopoly position on the upstream market must set its prices in a manner that allows its downstream competitors to remain on the market, whatever their efficiency. This position is at odds with US current case law according to which a *legitimate* monopolist has no duty towards its downstream competitors. However, the situation may be different in regulated industries for which the market positions are not only related to the merits. This is precisely the point stressed by the Supreme Court in *Trinko*: the EFD is no longer seen as an antitrust-related concept but only as remedy available to a sector specific regulator to organize the third party access to a *natural monopoly* network.

A Risk in Terms of Incentives According to the Economic Literature

Economic scholars have also harshly criticized the implementation of the EFD in the competition law field. Its broad implementation may lead to bind a successful firm to make its competitors benefit from its past investments, e.g., to free-ride them. Such a mandatory access may pertain to an asymmetric regulation of competition and it may have a deterring effect on its incentives to invest. *Ex post*, imposing an access to the incumbent's assets can be seen as an expropriation. The undertaking controlling the essential facility is deprived of the right to refuse the access and of its capacity to set freely its price. The economic gains resulting from its investment will be partially lost. *Ex ante*, anticipating such a risk might deter to invest in this type of asset. Depending on the level of the access charge, it may also impair its incentives to spend money in its maintenance or in its development, as the dominant undertaking cannot expect to extract all the economic gains resulting from its investment. In addition, it may even impair the incentives to invest for nondominant undertakings. As the new entrants may prefer to use the incumbent's assets, it may reduce the incentives to develop a new (essential) facility. Indeed, this case law may lead to expect the possibility to be bound to provide an access benefiting to competitors and in doing so to reduce the potential pay-offs associated to the initial investment.

As a consequence, imposing a mandatory access may be welfare-enhancing only in a static sense: it effectively increases the competition on the downstream market. However, in a dynamic sense, it may be a counterproductive remedy as it hinders the incentives to invest in the infrastructure and in strategies aiming at by-passing it. The consumer welfare gain resulting from the competitive remedy may be only a short-term one. Investments aiming at producing disruptive innovations may be discouraged and the asset owner may be daunted to invest to increase the quality of the service provided by its infrastructure or its capacities because it will also benefit to its competitors. In a worst-case scenario, the incumbent might be incentivized to under-invest to generate socially counter-productive bottlenecks in order to have an objective reason to deny the access to its competitors.

In addition, the very cautious approach of the economic literature considering the EFD implementation is even exacerbated for intangible assets as intellectual property rights (hereafter IP). For instance, how to characterize an intangible asset as absolutely necessary to a competitor to access the market and reasonably impossible to replicate in technical or in financial terms? Such preventions echo with the Stigler (1968) views according to which barriers to entry cannot really be technical or financial but are more commonly the product of public regulations. Moreover, public intervention through antitrust remedies are viewed with suspicion. It is particularly the case since these remedies are implemented within market building objectives (or in order to counterbalance individual market power) or since they aim at correcting the excess of IP rights protection. The risk of regulatory capture cannot be excluded.

Indeed, we may put into relief the long-standing mistrust of economic scholars towards antitrust suits initiated by competitors. The risk is to protect unduly competitors at the expense of consumers (Bork 1978). In addition, the incentives produced by competitive threats might be still possible despite a monopoly situation. The competitive pressure persists for as long as the

market remains a constable one. Compelling a monopoly to renounce to its contractual freedom and to limit its capacity to extract all the surplus created can eventually be prejudicial in terms of consumer welfare. In other words, compulsory licensing or mandatory access may lead to an asymmetric regulation of competition. If the EFD one may be acceptable as a regulatory tool in the framework of a sector specific liberalization, it is by far more debatable since the dominant undertaking's position is due to its merits and not to former exclusive rights.

A Broad Acclimatization in the EU Competition Law Enforcement

Even through the essential facility doctrine is now rejected as an antitrust tool in the USA, its implementation by the EU Commission and by the Member States' competition authorities is significant. The reason of this adherence may be explained by the EFD theoretical consistency with some of the EU competition policy underlying principles. Firstly, the EFD tackles the issues related to deny to access to the market (exclusionary abuse) or to excessive access prices (exploitative abuses). Secondly, as soon as the purpose of the competition law is conceived as encompassing the building competitive markets or the guarantee a level playing field, the EFD appears as an efficient tool to craft remedies to achieve such ends.

Implementation to Network Industries

The first uses of the EFD in the EU context naturally took place in the field of the liberalization of network industries with the 1992 EU Commission's decision in the case *B&I – Sealink* (Commission, case IV/34.174 – Sealink/B&I –Holyhead, 11 June 1992). It was also the case, 4 years later in the French case the 1996 competition authority's decision related to the Narbonne's heliport (French Competition Council, decision n°96-D-51, 3 September 1996). Through the EU Commission led liberalization process, the EFD was successively applied to telecommunications, electricity, gas, and railways sectors. We will see below that one additional special feature of the EU use of the EFD compared to the US case has to be taken into account: its large implementation to intangible assets. For instance, incumbents are commonly bound to share their consumer's data bases with new entrants in order ensure a level playing field (see for instance the French Competition Authority decision, n°17-D-06 related to practices in the markets of gas natural, electricity, and energy related services, 21 March 2017).

The EFD was often implemented as a competitive remedy in the sectors concerned by a liberalization process. The first step of liberalization commonly consists in guaranteeing an open (and regulated) access to the incumbent's infrastructure for the new entrants in order to favour the development of a services-based competition. The second step may be an unbundling (incumbent's vertical de-integration) or the implementation of the so-called *ladder of investment approach* (Bourreau et al. 2010).

The vertical unbundling is a drastic way to guarantee an undistorted access. It consists in the vertical integrated incumbent break up, as it was the case in the 1982 *AT&T* US case. These remedies were implemented in the energy sector exclusively through negotiated procedures (see for instance the *ENI* Commission decision 39.315, 29 September 2010).

The ladder of investment approach is a less intrusive intervention (without mandatory divestitures as structural remedies) that can be implemented since the competition is not potentially limited to the service but can be expanded to a part of the facilities. This approach consists in providing a broad access to the incumbent's facilities in the first steps of the liberalization process in order to lower barriers to entry at the benefits of its new competitors. The expected effect consists in helping them to acquire a sufficient customer's portfolio to overcome investors' risk aversion to finance their own infrastructures. This broad access will be progressively reduced in order to promote incentives to invest both for them and also for the incumbent, since the first ones cannot durably free ride its own investments. This approach allows to incentivize the new entrants to shift from a services-based competition to a facilities-based one. It does not concern the

natural monopoly in itself but all the facilities necessary to the connection with this one. Naturally regulated access to the natural monopoly segment remains at the end of this period. One of the interests of such a facilities-based competition is to favour a service differentiation among the competitors, for instance in terms of quality. It can be seen as broad way to implement the EFD within the context of a sector specific liberalization.

Even so, the case of the ladder of investment approach illustrates some of the common risks and pitfalls of the EFD implementation in network industries. On one hand, if the access charge is too low, the new entrants may free ride incumbent past investments and can be discouraged to invest in their own assets. On the other hand, the incumbent's incentives to develop and to maintain its own infrastructures may be reduced by such a mandatory access. Setting the access charge raises several difficulties. A high access charge level may erect a barrier to entry and unduly preserve inculbent's rents. Even if the access charge has to be oriented to its costs, it remains that information asymmetries might lead to significant regulatory costs and to possible suboptimal settings. In the same time, the access to the incumbent's facilities may be counterproductive for the new entrants as it might perpetuate a situation of dependence toward their main competitor. This dependence on the facilities controlled by their main competitor exposes them to the consequences of noncooperative behaviors as raising rival costs strategies (through the pricing of nonregulated ancillary services) or strategies aiming at artificially reducing the quality of the service they provide to final users.

The EU Commissions had faced sharp criticisms about its broad implementation of the EFD in the field of network industries. The concept of ladder of investments, presented above, illustrates that the scope of the essential facility can go far beyond the natural monopoly segment narrowly defined. The risk is to excessively extend the notion of essential facilities to all the specific and expensive assets that a new entrant needs to dispose to enter the market. Investing in some specific assets may actually represent a significant risk as their amortization can be difficult for a new entrant who only serves a small segment of the market. However, it is not so rather clear at the theoretical point of view that such investments can be considered as barriers to entry (see for instance Stigler 1968). The risk is to promote thought the EFD implementation an asymmetric regulation of the competition by mandating the incumbent to provide an access not only to its essential facilities but also to *convenient facilities* for its competitors (Ridyard 2004). The risk is to favor inefficient entries at the expense on final consumers.

A striking example of such concerns can be provided by the EU Commission *GVG* decision in 2003 related to an access of a German company to the Italian railways. The mandatory access covers all the components of the traction services encompassing the locomotives and the railway workers qualified for the Italian network themselves (European Commission, case COMP/37.685, GVG/FS, 27 August 2003). On the contrary, in the *Bronner* decision, the Court of Justice had decided that a mandatory access to an incumbent facility cannot be imposed since the new entrant can develop its own (Court of Justice, judgment C-7/97, Oscar Bronner GmbH & Co. KG v Mediaprint Zeitungs, 26 November 1996). An essential facility should be a nonreplicable one and not only a less expensive one. The EU case law has been overturned at the benefit of a broader definition of the notion of the essential facility.

Another spectacular example of the broad implementation of the EFD in the EU competition policy applied to the liberalisation of network industries can be provided by a French competition authority decision related to the electricity market in December 2007, *Direct Energy* (French Competition Council, decision n°07-D-43 related to EDF market practices, 10 December 2007). These far reaching remedies can be explained by the very specific context of the French liberalization characterized by the coexistence of regulated tariffs at the retail level and market prices at the wholesale one. The new entrants were bound to set their retail prices at this threshold to attract customers. In the same time, as they did not control their own generation

capacities, they must buy the electricity on the wholesale market. Since, the regulated retail price is set according the generation marginal cost of the plants, mainly nuclear, operated by the incumbent, they were potentially squeezed since the wholesale price might be higher than this last one. The competitive claim was grounded on this margin squeeze produced by the combined effect of the downstream (regulated) price and the upstream (market) price. The competition law-based remedy consisted in implementing the EFD to the energy generated by the nuclear power plants (de Hauteclocque et al. 2011). The new entrants benefit from a drawing right on the incumbent's nuclear plants originated generation. The EFD was implemented to energy itself through this quasi-structural remedy. Such types of remedies tend to ensure that the new entrants may be as efficient as the incumbent. The use of the EFD is not limited to guaranteeing an access to the market for the competitors of a vertically integrated incumbent but participates to a regulation of the competition. In other words, the access does not concern a network infrastructure but the upstream good. The purpose is to ensure a level playing field in the downstream market. . . even if it supposes to "subsidize" potentially less efficient competitors. The EFD in this context leads to horizontal wealth transfer among market players. In this instance, it was used to address a regulatory imperfection.

Implementation to Intellectual Property Rights

Beyond this broad implementation, the EU implementation of the EFD is also distinguishable from the US one, as it also broadly concerns intangible assets as intellectual property rights. The EFD was implemented in the EU case law on TV programs, data bases, and to interoperability devices between softwares. While the EFD as competition remedies may be admitted as the essential facility at stake resulted from exclusive rights, the case of intangible assets raises concerns as soon as the assets are the products of incumbent's merits. In the same time, these far reaching remedies can be analyzed as a mean to counterbalance the excessive protection of the incumbents produced by intellectual property rights. They can also be analyzed as a mean to prevent an unchallengeable monopolization of a given market or to counteract leveraging strategies. The leverage from the then near-monopoly Windows operating system was one of the main concerns 11 years ago (EU Court of First Instance, case T-201/04, Microsoft Corp. v. European Commission, 17 September 2007); (EU Commission, 18 July 2018, case 40099) and some online platforms raise nowadays the same types of competition issues.

In a historical perspective, the first implementation to the EFD to intangible assets in the case of the EU competition law took place in the 1990s with the Irish TV programs case (EU Court of Justice, cases C-241/91P and C-242/91P, RTE v European Commission, 6 April 1995). The dominant broadcasting operator edited its own TV programs magazine and refused to provide its TV listings to a new entrant who aimed at proposing a multi-channels TV programs magazine. The 1995 EU Court of Justice rulings imposed to communicate to this entrant the TV listings by considering that the refusal impair it to propose a new product to consumers. A second emblematic case was the IMS Health one (EU Court of Justice, case C-418/01, IMS Health Gmbh v NDC GmbH, 29 April 2004). An undertaking proposing software-based reporting tools for pharmaceutical sales data refused that a new entrant adopts the same database modular structure than its one, based on German Zip codes. The EU Court of Justice dismissed its claims in 2004 and considered that it must provide a licence for allowing its competitor to use the same database structure despite its protection through intellectual property rights. What is specific in this second case is that there is no new product at stake as it was the case in the 1996 ruling.

This point is all the more relevant that the initial EU case law imposed to fulfil several criteria to decide of a compulsory licensing remedy under the EFD. The 1995 Magill decision was for instance grounded on the future availability of a new product for consumers. In the IMS judgment, this condition was not required. The access may be required even if the new competitor will provide the same service in the same relevant

market. In the same way, in the Microsoft case, the Commission imposed as a remedy the sharing of the interface protocols with its OS, considering that the interoperability will favor the future development of new products. Such pragmatism was also at stake concerning another condition to implement the EFD for intangible assets as for instance the requirement to implement a balance of incentives before imposing a licensing. Indeed, the EU case law initially imposed to perform such a balance in order to assess the net effect on the incentives to invest and to innovate both for the incumbent and for the beneficiary of the compulsory licensing. Such an assessment, which echoes with the more economic approach of the competition law enforcement – e.g., its effects-based approach – is seldom, or almost never, implemented in the EU case law related to the EFD (Marty and Pillot 2012). In the EU Commission guidance related to its enforcement priority regarding the article 82 of the Treaty (now the article 102), issued in February 2009, only three criteria remain: the refusal must relate to a product or a service objectively necessary to effectively compete on the market; the refusal may lead to an elimination of the effective competition; and this refusal is likely to lead to consumer harm).

Discussion

Implementing the EFD for intangible assets raises specific concerns (Castaldo and Nicita 2007). Setting the licensing fee is one of the more delicate aspects. Defining the access charge for a physical network may already be a challengeable task. Even though the principles are clear (the access price should be oriented toward the cost), the influence of accounting choices and the imperfections of information may lead to significant difficulties. Moreover, if we consider the Direct Energy remedies how to set the real cost of an electricity generated by nuclear power plants? The solution applied in this specific case – an auction procedure – illustrates the difficulties encountered to implement such remedies. Things are even worse for intangible assets for which the cost are by far more difficult to assess. The access and replication costs are negligible. The initial investments and the risks taken cannot be easy to evaluate and may be irrelevant, as testifies the IMS case, in which the essential facility at stake was an IP right on the structure of data base in accordance with Zip codes.

In addition, a broad implementation of the EFD to intangibles might lead to price regulation and to some extend to competition regulation. Defining the terms of a FRAND licence terms (*Fair, Reasonable, and Non Discriminatory*) or balancing market powers among competitors impose to define the concepts of fairness and reasonableness. Furthermore, using the EFD to promote the liberalization of a network industry, or even worse to make a dominated market contestable, may lead the competition authority to act as a sector specific regulator aiming at favoring the development of the competitors, possibly at the expense of final consumers.

Finally, the EFD implementation as a competition law remedy illustrates the transatlantic divergences on several topics. First, it stresses the debates on the respective scopes of competition law enforcement and sector specific regulation. Second, it illustrates the divergences about the possibility to use competition remedies to address issues as the excessive protection granted by intellectual property laws. Third, it underlines the disagreements on the use of competition law remedies as tools to prevent or to correct excessive concentrations of economic powers, even if these ones mainly result from the merits.

However, a striking point in the analysis of the implementation of the EFD in the EU competition law enforcement should be put into relief. It consists in the sharp contrast between the frequency of its use and the reluctance of the competition authorities to ground decisions and to justify remedies explicitly on this doctrine. Several reasons might be highlighted. A first one may be found in its lack of well-grounded economic theory foundations, especially in the framework of an effects-based approach (Geradin 2004). A second one may be related to the relative fuzziness of the concept at legal point of view and to the possible risks induced in terms of decision reversal in the judicial control. As a consequence, the doctrine is

in the same time significantly used to shape competition law remedies in the EU case and rarely qualified as a legal basis to ground them.

Cross-References

► Negotiated Procedures in EU Competition Law

References

Areeda P (1990) essential facilities: an epithet in need of limiting principles. Antitrust Law J 58:841
Bork R (1978) The antitrust paradox. Free Press, New York
Bourreau M, Doan P, Manant M (2010) A critical review of the "ladder investment" approach. Telecommun Policy 34(11):683–696
Carlton D, Heyer K (2008) Assessing single-firm conduct. Comp Policy Int 4(2):285–305
Castaldo A, Nicita A (2007) Essential facility access in europe: building a test for antitrust policy. Rev Law Econ 3:83–110
de Hauteclocque A, Marty F, Pillot J (2011) The essential facilities doctrine in European competition policy: the case of the energy sector. In: Glachant J-M, Finon D, de Hauteclocque A (eds) Competition, contracts and electricity markets: a new perspective. Edward Elgar, London, pp 259–292
Geradin D (2004) Limiting the scope of Article 82 of the EC treaty: what can the EU learn from the US Supreme Court's Judgment in Trinko in the Wake of Microsoft, IMS, and Deutsche Telekom. Common Mark Law Rev 41:1519–1533
Marty F, Pillot J (2012) Intellectual property rights, interoperability and compulsory licensing: merits and limits of the European approach. J Innov Econ 9(1):35–61
Ridyard D (2004) Compulsory access under EU competition law: a new doctrine of "convenient facilities" and the case for price regulation. Eur Comp Law Rev 11:669–674
Stigler GJ (1968) The organization of industry. University of Chicago Press, Chicago

Ethical Review Board

► Institutional Review Board

EU Citizenship

► European Nationality

EU Microsoft Competition Case

Luca Rubini
University of Birmingham, Birmingham, UK

Abstract

The investigation and litigation that involved Microsoft before the EU authorities in the 2000–2012 period is arguably the first case that put the complex issues of the new economy to the test of EU competition laws. The case, which attracted a lot of attention, also beyond competition law circles, was a complex one, at various levels – technical, economic, and legal – and one that eventually interrogated the ultimate goals of competition policy in a given legal system. This entry takes the reader through the various issues and findings of the European Commission and the Court of First Instance of the European Union before offering few notes of analysis and a view on the future implications of this landmark case.

EL classification K4 Legal Procedure, the Legal System, and Illegal Behavior.

Background and Significance of the Case

"The Microsoft case was the monopolization case of the new economy just as *Standard Oil* had been the icon of the old" (Peritz 2010). This quote makes reference to the US case involving Microsoft and draws a momentous parallel with the famous *Standard Oil* case of 1911 where the US Supreme Court found Standard Oil guilty of monopolization of the petroleum industry and famously divided the company into several different and competing companies. This quote gives a very good idea of the importance of the competition law tangles that have involved Microsoft, the American new economy giant, in the USA. Equally important was the investigation and litigation that involved Microsoft in the EU in the 2000s and which produced some of the most

controversial decisions of the EU Commission and of the General Court.

Microsoft, founded by Bill Gates in the 1970s in California, has fast become one of the most successful companies of our time. It belongs to that inner circle of businesses strongly characterized by a unique blend of innovation and creativity and vision and aggressiveness – which can often be attributed to the leadership of unique founder individuals. Microsoft operates in some of the most crucial and innovative markets. Its products, and notably its operating systems and software applications, are ubiquitous. Their importance cannot be understated. They have influenced, indeed shaped, everyday life of billions of people around the world. Commercial success and especially market power attract the attention of antitrust control. Microsoft has thus been subject to various investigations and actions, based on antitrust laws, at both sides of the Atlantic. Since the beginning of the 1990s, Microsoft started to be the subject of the scrutiny of US authorities (on the US case, see Page and Lopatka 2007; Peritz 2010; Gavil and First 2014). This culminated in a much-publicized litigation at the turn of the millennium where it was charged with the allegedly illegal tying of the Internet Explorer web browser with the Windows operating system. Microsoft would have leveraged on the sheer dissemination of its PC operating system to extend its dominance in the web browser market. Following, among other things, a change in administration (from Clinton to Bush Jr.) and numerous settlements between the various parties involved, this litigation finished with no action. The focus shifted to the other side of the Atlantic. At the beginning of the 2000, the EU Commission officially started to investigate two conducts of Microsoft – the tying of its media player with Windows and the refusal of certain interoperability information to communicate with its server operating systems – which, in a momentous decision of 2004, were concluded to be breaches of EU antitrust law. In 2007, in a long-awaited decision, the General Court of the EU essentially confirmed the findings and rulings of the EU Commission. In 2009, the EU Commission started a new investigation focused (like the US case) on the bundling of Windows with Internet Explorer, which was settled the next year.

After summarizing the arguments and the findings of the EU case (section "The Decisions of the EU Commission and the General Court"), this entry provides few notes of analysis of the technical, economic, and legal aspects of this case (section "Analysis of the Case"). This analysis paves the way to the consideration of the broader implications of the case (section "Implications for the Future").

The Decisions of the EU Commission and the General Court

Article 102 of the Treaty on the Functioning of the EU (TFEU) (formerly, Article 82 of the EC Treaty) was the key provision at the center of both the EU Commission investigation and the litigation before the General Court. In the parts relevant to the Microsoft case, this provision, which dates back to the original Treaty of Rome of 1957, reads:

> Any abuse by one or more undertakings of a dominant position within the internal market or in a substantial part of it shall be prohibited as incompatible with the internal market in so far as it may affect trade between Member States.
>
> Such abuse may, in particular, consist in:
>
> . . .
>
> (b) limiting production, markets or technical development to the prejudice of consumers;
>
> . . .
>
> (d) making the conclusion of contracts subject to acceptance by the other parties of supplementary obligations which, by their nature or according to commercial usage, have no connection with the subject of such contracts.

The reader can find in the literature (Rubini 2010) a uniquely interesting repetition of the EU investigation and litigation, performed by its very same actors: the Commission (Banasevic and Hellström 2010), Microsoft (Forrester 2010; Bellis and Kasten 2010), and the General Court (Vesterdorf 2010). The following section summarizes the key arguments and findings of the case. This overview will introduce the reader to the various levels of complexity of this momentous case.

The Decision of the EU Commission

After a lengthy investigation, the EU Commission issued its decision on 24 March 2004. It concluded that Microsoft had violated Article 82 of the EC Treaty by abusing its dominant position in client PC operating system (OS) in two ways. Firstly, it had illegally refused to supply interoperability information which was indispensable for rival vendors to compete in the work group server operating system market. Secondly, it had illegally made the availability of the Windows client PC operating system conditional on the simultaneous acquisition of the Windows Media Player software. The Commission ordered Microsoft to disclose interoperability information and appointed a trustee to monitor the compliance with this duty. It also ordered Microsoft to provide a version of Windows without Windows Media Player. The Commission finally imposed a fine of 497,196 million euros for what it found to be a very serious infringement of Article 102 TFEU.

With respect to the **licensing** part of the case, the Commission found that Microsoft had breached Article 102 TFEU by refusing to supply interoperability information to its competitors. As noted, letter (b) of the same provision enlists: "limiting production, markets or technical development to the prejudice of consumers." Two Microsoft products, and markets, were under examination in the licensing claim: first, the Windows PC OS and, second, the Windows Server OS. The Commission found that, with market shares of, respectively, 93% and 60%, Microsoft was superdominant and dominant in the relevant product markets. The servers' market was the one where there had been, and there still was, some competition. Some companies were competing with Microsoft and were concerned by the reduction in the interoperability information available for communication between their servers and Windows PCs. The main issue of contention concerned the degree of interoperability, and amount of information, which should be guaranteed. The Commission opined that Microsoft refused information that was indispensable for external servers to interoperate with Windows, which limited innovation and risked eliminating competition to the detriment of consumers. In economic terms, the scenario was one of leveraging in the server OS market on the basis of the virtual monopoly held in the PC OS market. The significant network effects that are typical of these markets exacerbated the effects of this exclusionary conduct (Gil-Moltó 2010). The Commission therefore ordered Microsoft to license on reasonable and nondiscriminatory terms the relevant interoperability information to its competitors.

With respect to the **tying** claim, the Commission found that Microsoft had infringed Article 102 TFEU by bundling its Windows Media Player with its operating system. This had enabled Microsoft to expand its market power and foreclose competition in the media player market. The Commission based its decision on four steps: (i) the tying (i.e., Windows) and tied (i.e., Windows Media Player) products are separate products, (ii) Microsoft is dominant in the market for Windows, (iii) it does not give its customers a choice to obtain the tying product without the tied product, and (iv) this practice forecloses competition (in market of media players). The Commission thus ordered Microsoft to offer a version of Windows without the Windows Media Player.

The Judgment of the EU General Court

With a decision dated 17 September 2007, the General Court (at the time Court of First Instance) of the EU upheld the substance of the Commission's decision with respect to both abuses as well as the fine. The General Court's decision was not appealed before the Court of Justice of the EU.

The Duty to Disclose Interoperability Information

Before the General Court, Microsoft put forward various arguments to discredit the assessment of the EU Commission of the need for a duty to disclose interoperability information. Firstly, the concept of interoperability used by the Commission was too broad, and to follow it would have meant to enable Microsoft's competitors to have access to highly technologically innovative protocols and to clone its products. Secondly, the Commission did not follow the strict conditions that, according to the case law, need to be present before antitrust law can impose a duty to license

intellectual property. In particular, the presence of five alternative routes for interoperability defeated the existence of indispensability. In addition, a high threshold for antitrust intervention should be followed. The presence of competing servers in the market proved that competition was not being eliminated. Moreover, the emergence of no "new product" was prevented by Microsoft's conduct. Finally, Microsoft's refusal was justified on the grounds of protecting valuable intellectual property, and the forced disclosure of information would encourage copying and reduce incentives to innovate.

The General Court in turn rejected each of these arguments.

Firstly, the General Court confirmed the degree of interoperability the Commission found to be necessary to remain viable in the market and in the specific context of a Windows work group network (paras. 229–230). In particular, the Court found that the Commission was right in concluding that "the common ability to be part of [the Windows domain architecture] is a feature of compatibility between Windows client PCs and Windows work group servers" (para. 189). Consequently, the Commission was held to be correct in considering that the required interoperability information should cover "the complete and accurate specifications for all the protocols [that are] implemented in Windows work group server operating systems and that are used by Windows work group servers to deliver file and print services and group and user administration services, including the Windows domain controller services, Active Directory services and Group Policy services, to Windows work group networks" (para. 195). The Court also noted that this information does not extend "to the internal structure or to the source code of its products" (para. 206). The Court also rejected Microsoft's argument that this degree of interoperability would have allowed its competitors to clone its products or certain features of those products (paras. 234–242).

Secondly, the General Court also found the Commission did not err when it found that the information concerning the interoperability with the Windows domain architecture was **indispensable**. The availability of the required information was necessary in order to be able to compete viably with Windows work group server OSs on an equal footing.

On the one hand, the Court rejected that the five alternatives to disclosure could ensure the necessary degree of interoperability. Microsoft had put forward five methods which, though not ensuring the perfect substitutability that the Commission considers essential, made nonetheless it possible to achieve the "minimum level of interoperability required for effective competition," "to work well together" (paras. 345–346). It is interesting to note here that, in its Decision, the Commission had already rejected "reverse engineering" as a viable alternative to disclosure of interoperability information (see para 562 of the Decision).

On the other hand, on the issue of what level of **elimination of competition** is necessary to trigger antitrust intervention, Microsoft contended that the Commission had been satisfied with establishing a mere "risk" of the elimination of competition in the work group server OS market. What should have been required was the "likelihood" or, in other words, the "high probability" of distorting competition. The General Court responded that Microsoft's complaint was "purely one of terminology" and "wholly irrelevant" (para. 561). The Court went on noting,

> The expressions 'risk of elimination of competition' and 'likely to eliminate competition' are used without distinction by the Community judicature to reflect the same idea, namely that Article [102 TFEU] does not apply only from the time when there is no more, or practically no more, competition on the market. If the Commission were required to wait until competitors were eliminated from the market, or until their elimination was sufficiently imminent, before being able to take action under Article [102 TFEU], that would clearly run counter to the objective of that provision, which is to maintain undistorted competition in the common market and, in particular, to safeguard the competition that still exists on the relevant market.
>
> In this case, the Commission had all the more reason to apply Article [102 TFEU] before the elimination of competition on the work group server operating systems market had become a reality because that market is characterized by significant network effects and because the elimination of competition would therefore be difficult to reverse. . . .

> Nor is it necessary to demonstrate that all competition on the market would be eliminated. What matters, for the purposes of establishing an infringement of Article [102 TFEU], is that the refusal at issue is liable to, or is likely to, eliminate all effective competition on the market. It must be made clear that the fact that the competitors of the dominant undertaking retain a marginal presence in certain niches on the market cannot suffice to substantiate the existence of such competition (paras. 561–563).

Applying this reasoning to the actual market data, the General Court concluded that "Microsoft's refusal has the consequence that its competitors' products are confined to marginal positions or even made unprofitable. The fact that there may be marginal competition between operators on the market cannot therefore invalidate the Commission's argument that all effective competition was at risk of being eliminated on that market" (para. 593).

Citing the previous case law, Microsoft argued that it has not been established that its refusal prevented the appearance of a "new product" for which there is unsatisfied consumer demand. The General Court noted that the appearance of a new product "cannot be the only parameter which determines whether a refusal to license an intellectual property right is capable of causing prejudice to consumers" (para. 647). Such prejudice may arise, as the Commission had found, also when there is a limitation of technical development, as provided in Article 102 TFEU, letter (b) (ibid).

The General Court then endorsed the Commission's finding that "the information at issue does not extend to implementation details or to other features of Microsoft's source code," representing "only a minimum part of the entire set of protocols implemented in Windows work group server operating systems" (paras. 657–658). Nor – the Court noted – "would Microsoft's competitors have any interest in merely reproducing Windows work group server operating systems." Once the necessary interoperability information is made available, "they will have no other choice, if they wish to take advantage of a competitive advantage over Microsoft and maintain a profitable presence on the market, than to differentiate their products from Microsoft's products" (para. 658).

Finally, the General Court noted that the mere fact that the relevant technology is covered by intellectual property protection is not per se a sufficient objective justification (para. 690). Microsoft had not sufficiently established that, if it were required to disclose the information, disclosure would have a significant negative impact on its incentives to innovate (para. 701).

E

The Bundling of the Windows Media Player

As regards the tying claim, the analysis immediately focused on whether the Commission introduced new law in the area by examining whether the dominant undertaking "does not give customers a choice to obtain the tying product without the tied product" (para. 845). The General Court dismissed Microsoft claim that this was any different from what Article 102 EC (d) requires, i.e., that bundling assumes that consumers are compelled, directly or indirectly, to accept supplementary obligations, such as those referred to in Article 102(d) EC (para. 864). The Court points out that "coercion is mainly applied first of all to OEMs [Original Equipment Manufacturers], who then pass it on to the end user" (para. 865). The analysis then shifted to the question whether the Commission introduced a new condition, analyzing whether Microsoft's conduct foreclosed competition. Microsoft accused the Commission of having added a new test, nor provided for in the law (and in particular Article 102 TFEU, letter d), notably that the tying conduct is foreclosing competition (para. 846). The Court acknowledged that, in light of the specific circumstances of the case, the Commission did not "merely assume, as it normally does in cases of abusive tying, that the tying of a specific dominant product has by its nature a foreclosure effect" (para. 868). By contrast, it "examined more closely the actual effects which the bundling had already had on the streaming media player market and also the way in which that market was likely to evolve" (ibid). The Commission was right in doing this, since the "list of abusive practices set out in the second paragraph of Article 82 EC is not exhaustive and that the practices

mentioned there are merely examples of abuse of a dominant position" (para. 860). More specifically, bundling by an undertaking in a dominant position may also infringe Article 82 EC where it does not correspond to the example given in Article 82(d) EC. Accordingly, in order to establish the existence of abusive bundling, the Commission was correct to rely in the contested decision on Article 82 EC in its entirety and not exclusively on Article 82(d) EC (para. 861). The Court, in any event, concluded that constituent elements of abusive tying identified by the Commission coincide effectively with the conditions laid down in Article 82(d) EC (para. 862).

The second point of contention focused on whether media functionality is a separate product from the PC operating system, which is a necessary condition to have tying. Microsoft argued this was not the case (para. 912). Media functionality forms an integral part of the operating system with the consequence that what is at issue is a single product. This would have been confirmed by customers' expectation that any PC OS have essential audio and video functionalities. While recognizing the rapid evolution of the industry, where products that are initially separate are then subsequently regarded as forming a single product, the Court concluded that the Commission was correct to find two separate products in the period of the investigation (paras. 913–914). The Court noted that there exists a separate consumer demand for streaming media players (para. 917). The simple fact that two separate products are complementary does not exclude their difference (paras. 921–922). The Court confirmed the Commission's finding that media functionality is not linked, by nature or by commercial usage, to PC OSs (para. 938) Firstly, it does not seem that the two constitute by nature indissociable products (para. 939). In any event, the Court importantly underlines that "it is settled case-law that even when the tying of two products is consistent with commercial usage or when there is a natural link between the two products in question, it may none the less constitute abuse within the meaning of Article 82 EC, unless it is objectively justified" (para. 942). Secondly, "it is difficult to speak of commercial usage in an industry that is 95% controlled by Microsoft" (para. 940). Furthermore, the fact that vendors of competing client PC OSs also bundle those systems with a media player was not conclusive (para. 941). The Court crucially noted that "it is settled case-law that even when the tying of two products is consistent with commercial usage or when there is a natural link between the two products in question, it may none the less constitute abuse within the meaning of Article [102 TFEU], unless it is objectively justified" (para. 942). Microsoft's argument that the integration of Windows Media Player in the OS was dictated by technical reasons was found not to be substantiated (para. 937).

There was no real contention with respect to Microsoft's dominance in the PC OS market (para. 870) or with respect to the inability of consumers to obtain Windows without Windows Media Player (para. 961).

The Court then moved to the actual assessment of how Microsoft's conduct would have foreclosed competition; the General Court upheld the Commission's analysis.

Microsoft contended that the Commission, recognizing that it was not dealing with a classical tying case, had to apply a new and highly speculative theory, relying on a prospective analysis of the possible reactions of third parties, in order to reach the conclusion that the tying at issue was likely to foreclose competition. The General Court noted that "it is clear that, owing to the bundling, Windows Media Player enjoyed an unparalleled presence on clients PCs throughout the world, because it automatically achieved a level of market penetration corresponding to that of the Windows client PC operating system" (para. 1038). The pre-installation of Windows Media Player made users less likely to use alternative players (para. 1041). The said bundling created disincentives for OEMs to ship third-party media players on their client PCs for technical (e.g., higher usage of hard-disk space, risk of confusion on the part of users, increase of customer support and testing costs) and economic (e.g., higher price of the PC) reasons (paras. 1043–1045). The Court also found that the Commission was also correct to find that methods of distributing media players other than

pre-installation by OEMs could not offset Windows Media Player's ubiquity (paras. 1049–1057). The Court confirmed the Commission's finding that the market for streaming media player was characterized by significant indirect network effects (paras. 1061–1076). That expression describes the phenomenon where the greater the number of users of a given software platform, the more there will be invested in developing products compatible with that platform, which, in turn, reinforces the popularity of that platform with users (para. 1061).

Finally, Microsoft's arguments that the integration of media functionality in Windows was "indispensable in order for software developers and internet site creators to be able to continue to benefit from the significant advantages offered by the "stable and well-defined" Windows platform" were rejected (para. 1146). The Court further noted that, although standardization may be positive (but not necessarily wanted by third parties), it cannot be allowed to be imposed unilaterally by a dominant firm (paras. 1152–1153). Similarly, the claim that the Commission was interfering with Microsoft's business model was not accepted (paras. 1149–1150). The Commission did not deny that Microsoft could provide a version of Windows with Windows Media Player. It objected to the fact that this is the only available version and, more specifically, that no version of the ubiquitous OS is offered without Windows Media Player.

Analysis of the Case

The summary exposition of section II gives a good flavor of the complexity of this case. Indeed, it is various factors of technical, economic, and legal nature, each adding to the other, that make it complex. In turn, this complexity explains how the assessment of both the EU Commission and the General Court could be so controversial and generate much debate. Quite probably, the most interesting remarks in this discussion come from Bo Vesterdorf, the President of the General Court in the very same case, who, focusing on the duty to disclose interoperability information, expressed concerns about the possible negative impact of the General Court's decision on investment and innovation (Vesterdorf 2008).

The first layer of complexity comes from the highly **technical** nature of the subject matter. What clearly comes out from an overview of the literature is that, without a proper understanding of the "basic technology" issues, any economic or legal assessment is doomed to fail (see the excellent primer of Jackson 2010). The "technical basis" of the EC case "was significantly more complex than its US counterpart," but, unfortunately, "many of the articles written about the Microsoft proceedings fail to convey the technical complexity of the issues at stake" (Jackson 2010). It is sometimes (wrongly) assumed that the computing sector is just like any other industry and that a stylized account of the technical facts would suffice. It should also be highlighted that it is in particular the "interoperability" part of the case that requires an unusually detailed level of technical understanding, which probably explains why this claim attracted particular attention. To name just few of the technical issues of the case and indeed the most general ones: What is interoperability? In particular, what degree of interoperability is necessary for computer programs to be meaningfully interfaced? Consequently, what type of information is necessary? Apart from disclosure, what paths are available to software developers to obtain this information? What difficulties do they each involve? How effective are they to ensure interoperability? What does define a set of instructions as a computer program? Is it possible to distinguish the latter from a mere functionality? When do separate computer programs stop being so and become one single program? Can functionalities be removed from a computer program without impairing the integrity of the program? (for a discussion of these questions, see Walsh 2010; Andreangeli 2010).

The second level of complexity, which is directly based on the first one, concerns the **economics** of this sector and of the behavior of consumers and competitors. Some unique features characterize the computer industry, such as its very fast pace of development and innovation and the high relevance of network effects (see

Gil-Moltó 2010; Walsh 2010; Liebowitz and Margolis 2001; Evans et al. 2000). Thus, on the one hand, the tendency toward network effects makes foreclosure more likely. On the other hand, the fast development of the industry where today's winner is tomorrow's loser (and vice versa) represents a crucial concern for legal and regulatory processes, in two respects. First, the speed and uncertainty of change may make the assessment of harm and efficiencies difficult. Second, the time required for investigations may render legal processes continually and, inherently, obsolete. Equally, in such a dynamic environment, remedies are difficult to be tailored and effective (Economides and Lianos 2010).

The third, and final, level of complexity lies in the **law**. After almost 60 years from its entry into force, the provision at issue, Article 102 TFEU, is still vague. "Despite a significant volume of case law expanding, refining, clarifying and apparently applying the law, there is still considerable uncertainty about the exact conditions amounting to abuse" (Walsh 2010). The fact is that Article 102 TFEU, with its broad and general language, is a "standard" (rather than a "rule"), which means that the regulator or judge has to define its content on a case-by-case basis (see Kaplow 1992). This process may lead to more accurate results but is certainly more costly and uncertain. This uncertainly is particularly acute in technological sectors like those of the new economy and does significantly depend on the ambiguous signals that come from the first two layers of complexity set out above.

More deeply, the root of the uncertainty of Article 102 TFEU depends on the fact that its meaning changes depending on the different conceptions of what competition and antitrust law should be about, conceptions that succeed and prevail one after the other. In 60 years (and quite probably in many years to come), the language is (and will be) the same, but the understanding of competition policy and of the regulation of unilateral behavior by dominant firms has changed (and will change). What is competition law about? Is it (only) about efficiency? What do we mean by efficiency? Is competition law more about safeguarding a competitive process, characterized by market access and contestability, or, more simply, an efficient outcome of the market contest? What other competition policy objectives play within the frame of competition law? (For commentary around these questions, see Kerber 2008 and Fox 2008.)

That being said, one has to ask whether the Commission and the General Court have really been revolutionary in their legal analysis. We focus here on the interoperability issue. In particular, did they really depart so dramatically from the previous *Magill* and *IMS Health* case law and their formulation of the "exceptional circumstances" under which a duty to deal and supply can be enforced? Fox (2010) rightly observes that "[t]he Microsoft facts did not fit the factors very snugly; but they fit the concept of essentiality much better than the facts of either *IMS Health* or *Magill*". While the General Court (and the Commission) may have been essentially right, Fox goes on noting that it would have been "much more satisfying" if the General Court had more openly and directly asked the following questions:

(1) Are consumers and the market seriously disadvantaged by denial of full access to interoperability information? If the answer is "yes":
(2) Would the respondent and the market be seriously disadvantaged by a duty to grant access?

According to the facts on file, it seems that first answer would indeed have been positive (in the Court's decision, it is repeatedly mentioned that users preferred certain rivals' products on all qualities except interoperability) and the second one negative (it should again be remembered that, according to the facts, before achieving a significant presence in the market of server OSs, Microsoft provided complete interoperability information).

A similar exercise – but with probably a more uncertain outcome – could be carried out for bundling (see Andreangeli 2010). In brief, one could ask whether, through the bundling of the Windows Media Player with Windows, consumers

were really "coerced" to use Windows Media Player and, as a result, consumers (and competitors) were harmed by this conduct.

The relativity of competition conceptions, and the ensuing flexibility of antitrust laws, and in particular the regulation of unilateral behavior (the most flexible of them all), becomes especially apparent if a parallel is made between how the EU dealt with the conduct of Microsoft and how the US treated it – or, better, would treat it. This is particularly apparent with respect to the interoperability issue and to the imposition, via antitrust laws, of the duty to supply. As Fox (2010) again shows:

> Any analyst applying the law and spirit of *Trinko* [landmark decision where the US Supreme Court emphatically underlined that there is no duty to deal in American antitrust law] would not start the analysis with the question posed above: Are consumers seriously disadvantaged by work group suppliers' lack of seamless access to the standard operating-system network? Analysis would start with quite a different question: Why should Microsoft be ordered to share its property with anyone, let alone rivals? US courts generally presume that a duty to deal will seriously impair a monopoly firm's incentives – to the harm of the market and innovation.

A similar conclusion on the EU-US divide can be taken for the claim of illegal bundling. As shown by the Court of Appeals of DC in the US Microsoft case, US Courts do require a rule or reason assessment and, in so doing, set the threshold pretty high before concluding that bundling is anticompetitive in a new economy setting (see Andreangeli 2010).

Implications for the Future

Many have highlighted the true exceptionality of the facts of the Microsoft case, which would suggest that it is difficult to draw any lessons for the future. Still, we believe, two main implications can be detected.

Microsoft was the first in a new string of cases dealing with the new economy and forcing antitrust laws, which in decades had developed with more traditional industries, to be confronted with extremely difficult questions. More than ever economic and technical knowledge are essential. Current (at the time of writing) investigations like *Google* raise important questions on the definition of the relevant markets and of abuse in highly complex and interconnected market (see Pollock 2010; Lianos and Motchenkova 2012; Bork and Sidak 2012).

Secondly, it has been noted above that the EU has taken a "typically European" approach, which in the end is definitely more interventionist as compared to the US one. One may wonder whether this will continue in other new economy cases or whether a rapprochement with what happens across the Atlantic can be expected.

References

Andreangeli A (2010) Tying, technological integration and Article 82 EC: where do we go after the Microsoft case? In: Rubini L (2010), pp 318–343

Banasevic N, Hellström P (2010) Windows into the world of abuse of dominance: an analysis of the Commission's 2004 Microsoft Decision and the CFI's 2007 judgment'. In: Rubini L (2010), pp 47–75

Bellis JF, Kasten T (2010) The Microsoft Windows Media Player tying case. In: Rubini L (2010), pp 127–165

Bork RH, Sidak JG (2012) What does the Chicago School teach about Internet search and the antitrust treatment of Google. J Comp Law Econ 8(4):633–700

Economides N, Lianos I (2010) The quest for appropriate remedies in the EC Microsoft cases: a comparative appraisal. In: Rubini L (2010), pp 393–462

Evans DS, Fisher FM, Rubinfeld DL, Schmalensee RL (2000) Did microsoft harm consumers? Two opposing views. AEI-Brookings Joint Center for Regulatory Studies, Washington, DC

Forrester IS (2010) Victa placet mihi causa: the compulsory licensing part of the Microsoft case In: Rubini L (2010), pp 76–126

Fox EM (2008) The efficiency paradox. In: Pitofsky (ed) How the Chicago School overshot the mark: the effect of conservative economic analysis on US antitrust. Oxford University Press, New York, pp 77–100

Fox EM (2010) The EC Microsoft case and the duty to deal: the transatlantic divide. In: Rubini L (2010), pp 274–281

Gavil AI, First H (2014) The Microsoft antitrust cases. Competition policy for the twenty-first century. MIT Press, Cambridge, MA

Gil-Moltó MJ (2010) Economic aspects of the Microsoft case: networks, interoperability and competition. In: Rubini L (2010), pp 344–368

Jackson C (2010) The basic technology issues at stake. In: Rubini L (2010), pp 3–46
Kaplow L (1992) Rules versus standards: an economic analysis. Duke Law J 42:557–629
Kerber W (2008) Should competition law promote efficiency? Some reflections of an economist on the economic foundations of competition law. In: Drexl J, Idot L, Moneger J (eds) Economic theory and competition law. Edward Elgar, Cheltenham/Northampton, pp 93–120
Lianos I, Motchenkova E (2012) Market dominance and quality of search results in the search engine market, TILEC discussion paper DP 2012/036, 31 October 2012
Liebowitz SJ, Margolis SE (2001) Winners, losers & Microsoft. Competition and antitrust in high technology. The Independent Institute, Oakland
Page WH, Lopatka J (2007) The Microsoft case. Chicago University Press, Chicago
Peritz JRR (2010) The Microsoft chronicles. In: Rubini L (2010), pp 205–257
Pollock R (2010) Is Google the next Microsoft? Competition, welfare and regulation in online search. Rev Netw Econ 9(4):1–29
Rubini L (2010) Microsoft on trial. Legal and economic analysis of a transatlantic antitrust case. Edward Elgar Publishing, Northampton
Vesterdorf B (2008) Article 82 EC: where do we stand after the *Microsoft* judgment? Glob Antitrust Rev 1–41
Vesterdorf B (2010) Epilogue. In: Rubini L (2010), pp 487–489
Walsh A (2010) *Microsoft v Commission:* interoperability, emerging standards and innovation in the software industry. In: Rubini L (2010), pp 282–317

Eucken, Walter

Stefan Kolev
Wilhelm Röpke Institute, Erfurt and University of Applied Sciences Zwickau, Zwickau, Germany

Abstract

The purpose of this entry is to delineate the political economy of Walter Eucken. To reach this goal, a history of economics approach is harnessed. First, the entry concisely reconstructs Eucken's life, intellectual evolution, and heritage. Second, it presents the specificities of his "theory of orders" and his "order-based policy" gateway to a rule-based political economy.

Biography

Walter Eucken (1891–1950) was doubtlessly one of Germany's most important twentieth-century economists, especially in the field of political economy. The Freiburg School, which he co-initiated in the 1930s as an interdisciplinary group of economists and legal scholars, has had a crucial impact on the evolution of postwar German economics and legal scholarship, but also on the practical trajectory in the economic policies of the Federal Republic and of European integration. This introduction aims at embedding Eucken in his time and at depicting his role in several contexts in science as well as at the interface between science and society, before subsequently turning to his contributions to political economy.

Eucken was born in Jena into the family of the philosopher and later Nobel Prize laureate Rudolf Eucken. He studied a combination of history, economics, law, and administrative science at Kiel, Bonn, and Jena and was heavily influenced in his socialization in economics by the Younger Historical School, even though his teachers were not totally hostile to theorizing (Goldschmidt 2013, pp. 127–129). Having completed a dissertation (at Bonn) and a habilitation (at Berlin) on historicist grounds, he received a call to Tübingen in 1925 and subsequently to Freiburg in 1927, where he remained for the rest of his life. What would later become famous as the Freiburg School of Ordoliberalism came into existence in the early 1930s, when Eucken formed a scholarly community with the law professors Franz Böhm (1895–1977) and Hans Großmann-Doerth (1894–1944), also attracting talented younger scholars like Friedrich Lutz (1901–1975) and Leonhard Miksch (1901–1950) to Freiburg (Goldschmidt and Wohlgemuth 2008a). The cooperation of Eucken with Böhm and Großmann-Doerth steadily intensified, with the book series "Order of the Economy" initiated in 1936 as a milestone – its introduction under the title "Our Mission" became the programmatic manifesto of the incipient ordoliberal understanding of the role of law and economics in science and in society (Böhm et al. 2008; Goldschmidt

and Wohlgemuth 2008c). In the adverse intellectual climate of the time, Eucken and Böhm actively rejected National Socialism in theory and in practice, with Eucken openly opposing Martin Heidegger's rectorate policies in the university senate in 1933 (Nicholls 1994, pp. 60–67; Klinckowstroem 2000, pp. 85–88). Eucken was also among the very few who remained openly loyal to Edmund Husserl until his death in 1938, with his own epistemology heavily influenced by Husserl's phenomenology. Also, in the late 1930s and early 1940s, Eucken was an active participant in intellectual resistance groups, later to become known as the Freiburg Circles (Glossner 2010, pp. 31–38). Despite the isolation particularly during the war, Eucken remained connected to intellectual allies like F.A. Hayek and Wilhelm Röpke and in the immediate postwar years became a seminal figure in the international revitalization of liberalism, among others during the founding years of the Mont Pèlerin Society (Kolev et al. 2014). Simultaneously, Eucken was one of the principal policy advisors to the allies in Germany and also to Ludwig Erhard's increasingly influential policy strategy, later to become famous as the Social Market Economy (Goldschmidt and Wohlgemuth 2008b, pp. 262–264; White 2012, pp. 233–238). Eucken passed away in March 1950 while giving a lecture series at the London School of Economics upon Hayek's invitation.

While the relevance of ordoliberalism for postwar Germany and Europe is discussed in the entry dedicated to ordoliberalism, the impact of Eucken's heritage as a person deserves special attention. In 1954, the Walter Eucken Institute was founded by friends, colleagues, and students in the house of his family in Freiburg and is until today an influential research institute and think tank. Eucken's personality still attracts paramount attention: in 2014 Federal President Gauck delivered an address in Freiburg upon the 60th anniversary of the Walter Eucken Institut, in 2016 Chancellor Merkel delivered an address in Freiburg upon Eucken's 125th birthday, while today's economists argue about "Walter Eucken's long shadow" in the German policy responses to the Euro crisis (Feld et al. 2015; Bofinger 2016) and his relevance for related rule-based research programs like Constitutional Political Economy (Vanberg 1988; Buchanan 2012; Köhler and Kolev 2013).

Order, Science, and Policy

As described above, Eucken lived in a time as hostile as possible to liberty – and in a time which was characterized by extreme degrees of arbitrariness and chaos. This is one of the reasons why Eucken and his associates focused their political economy and social philosophy on the concept of "order" – a term which is among the most complex and most multifaceted in intellectual history (Anter 2007, pp. 127–158). Eucken's specific research program took shape relatively late in his life – it "matured slowly" (Hayek 1951, p. 337) – and can be interpreted as concentrating upon two central goals: first to enable a higher degree of order for designing economic theory and second to enable a higher degree of order for conducting economy policy. These two goals are also the focal points of two key terms in Eucken's system and of his two major books: *Ordnungstheorie* as his contribution to economic theory, presented 1940 in *The Foundations of Economics* (1992), and *Ordnungspolitik* as his contribution to economic policy, presented 1952 in his posthumous *Principles of Economic Policy* (2004). Important to underscore and to show below, both Eucken's theory and his perspective on economic policy are very much in line with the "problem of constitutional choice, i.e., as a question of how desirable economic order can be generated by creating an appropriate economic constitution" (Vanberg 2001, p. 40). The two domains of theory and policy offer a helpful structure for continuing this exposition.

Theory of Orders: An Alternative to the Ruins of Historicism and Pure Abstraction

German economics has been famous, and at times notorious, for its inclination toward extensive methodological and epistemological debates.

The notoriety stems from periods like the Weimar Republic, when many economists continued fighting about methods and epistemology without being able (or willing) to tackle the pressing problems of economic life (Köster 2011, pp. 41–60). Eucken's project was targeted at the very opposite: with his "Foundations" he attempted to identify a solid methodological and epistemological basis for theorizing not as an aim in itself, but rather to enable such kind of theorizing which can alleviate the immense problems of the age after National Socialism. This section focuses on those theoretical concepts that have direct practical impact and are thus indispensable for understanding Eucken's political economy.

First and foremost, the separation between "order" and "process" is crucial: "economic order" for Eucken is to be understood as the sum of market forms and monetary systems which frame the interactions of the individuals, whereas "economic process" stands for the interactions themselves. Synonymous for the framework of the economic order are "the rules of the game," whereas the economic process can be translated as "the moves of the game" (Eucken 1992, pp. 223–232, 2004, p. 54). An additional cornerstone of Eucken's system is "the interdependence of orders" – a concept emphasizing the embeddedness of the economic order within the other social orders, notably the law and the political order of the state. This is of special relevance since it shows that even though the different social orders have their own individual logics, these orders have to be thought in their diverse interrelations. Also, the combinability of economic, legal, and political orders is not arbitrary, i.e., specific combinations like a centrally planned economy and rule of law are hardly stable in the long term. At the same time, he underscored that having a well-developed rule of law is not a sufficient condition for a well-ordered market economy – rather, it is through the cooperation of economists and legal scholars that specific principles of the rule of law will be identified as indispensable preconditions and prerequisites for the market economy (Eucken 1992, pp. 85–90). Eucken's theory of orders thus made it possible to do away with the ruins of atheoretical or even antitheoretical historicism left by the Younger and the Youngest Historical Schools well into the 1930s, but at the same time he warned that abstract theorizing can be dangerous if its results are applied to any orders without carefully considering their specificities of time and space (Eucken 1992, pp. 41–44).

Order-Based Policy: An Alternative to Laissez-Faire, Central Planning, and Interventionism

The careful and systematic shaping of the economic order (synonymously: of the economic constitution) is at the core of Eucken's political economy. The above theoretical distinction of "order" and "process" enabled him to coin his famous *Ordnungspolitik*. A term difficult to translate precisely into any other language, it has been translated into English with the general term "rule-based policy" or with the more specific one "order-based policy." Its central target is a clear-cut conceptual alternative to laissez-faire, to the centrally planned economy, and to interventionism, using two criteria which Eucken distilled as essential for judging the merits of economic orders: the material criterion if an order is productive (i.e., overcoming scarcity as much as possible), and the ideal criterion if an order is humane (i.e., enabling a life in self-determination for the individuals) (Eucken 1992, pp. 239–241). Neither a laissez-faire regime nor a centrally planned economy fulfill these two criteria – a laissez-faire regime is deficient in terms of the framework of its economic order (with the implied belief in its automatic self-generation), while a centrally planned economy is deficient in terms of the properties of its economic process (and the impossibility of a self-determined life in it) (Eucken 1948). Eucken's "order-based policy" aims at optimally setting the rule-based framework of its economic order by a strong state envisioned as a referee impartial vis-à-vis vested interests, a state which (unlike the arbitrariness of interventionism) abstains from interventions into the economic

process except in well-defined exceptions (Giersch et al. 1992, pp. 28–32). Such a policy is aimed at fighting the first and foremost evil of social life: according to the Freiburg School, this evil is always the phenomenon of power in all forms of power relations – stemming from the state, from the market, and from other social orders (Eucken 2004, pp. 175–179; Foucault 2008, pp. 129–158). As first presented by Böhm and discussed in the entry covering his work, the ordoliberal instrument against power is competition. Eucken's contribution here is to specify the kind of principles a "competitive order" – the order conforming both to the "productive" and to the "humane" criterion – has to follow (Oliver 1960, pp. 133–140). He devised two sets of principles, which are at the core of this "competitive order": the "constitutive" and "regulating" principles. The former comprise a functioning price system, a sound currency, open markets, private property, freedom of contract, liability, and constancy of economic policy, while the latter aim at additionally curbing monopoly power, income inequalities, externalities, and problems of the labor market (Sally 1998, pp. 111–114; Kolev 2015, pp. 428–430).

Many of these problems may seem almost trivial today, but at the time of Eucken's death, they were anything but trivial – not in Europe and not in the United States, as the parallels between Eucken's endeavor and the project of the "Old Chicago" School clearly indicate (Van Horn 2009; Buchanan 2012; Köhler and Kolev 2013). Other problems, especially the issues of power in the digital age, the lack of liability in large sections of the political order and the financial system, or the search for maxims how to avoid the arbitrariness of interventionism and instead to base economic policy on rules, have lost nothing of their relevance.

Cross-References

- ▶ Austrian School of Economics
- ▶ Böhm, Franz
- ▶ Constitutional Political Economy
- ▶ Hayek, Friedrich August von
- ▶ Mises, Ludwig von
- ▶ Ordoliberalism
- ▶ Political Economy
- ▶ Röpke, Wilhelm
- ▶ Schmoller, Gustav von

References

Anter A (2007) Die Macht der Ordnung. Aspekte einer Grundkategorie des Politischen, 2nd edn. Mohr Siebeck, Tübingen

Bofinger P (2016) German macroeconomics: the long shadow of Walter Eucken. CEPR's Policy Portal of June 7 2016. Available at: http://voxeu.org/article/german-macroeconomics-long-shadow-walter-eucken

Böhm F, Eucken E, Großmann-Doerth H (2008) Unsere Aufgabe. In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger Tradition der Ordnungsökonomik. Mohr Siebeck, Tübingen, p 27–37. English translation In: Peacock A, Willgerodt H (eds) (1989) Germany's social market economy: origins and evolution. St. Martin's Press, New York, p 15–26

Buchanan JM (2012) Presentations 2010–2012 at the Jepson School of Leadership, Summer Institute for the History of Economic Thought, University of Richmond. Available at: http://jepson.richmond.edu/conferences/summer-institute/index.html

Eucken W (1948) On the theory of the centrally administered economy. An analysis of the German experiment, parts I and II. Economica 15(58):79–100. 15(59): 173–193

Eucken W (1992) The foundations of economics. History and theory in the analysis of economic reality, 2nd edn. Springer, Berlin

Eucken W (2004) Grundsätze der Wirtschaftspolitik, 7th edn. Mohr Siebeck, Tübingen

Feld LP, Köhler EA, Nientiedt D (2015) Ordoliberalism, pragmatism and the eurozone crisis: how the German tradition shaped economic policy in Europe. Eur Rev Int Stud 2(3):48–61

Foucault M (2008) The birth of biopolitics. Lectures at the Collège de France 1978–1979. Palgrave Macmillan, New York

Giersch H, Paqué K-H, Schmieding H (1992) The fading miracle. Four decades of market economy in Germany. Cambridge University Press, Cambridge

Glossner CL (2010) The making of the German post-war economy. Political communication and public reception of the social market economy after world war II. Tauris Academic Studies, London

Goldschmidt N (2013) Walter Eucken's place in the history of ideas. Rev Austrian Econ 26(3): 127–147

Goldschmidt N, Wohlgemuth M (2008a) Entstehung und Vermächtnis der Freiburger Tradition der Ordnungsökonomik. In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger Tradition der Ordnungsökonomik. Mohr Siebeck, Tübingen, pp 1–16
Goldschmidt N, Wohlgemuth M (2008b) Social market economy: origins, meanings, interpretations. Constit Polit Econ 19(3):261–276
Goldschmidt N, Wohlgemuth M (2008c) Zur Einführung: Unsere Aufgabe (1936). In: Goldschmidt N, Wohlgemuth M (eds) Grundtexte zur Freiburger Tradition der Ordnungsökonomik. Schweizer Monatshefte, Tübingen, pp 21–25
Hayek FA (1951) Die Überlieferung der Ideale der Wirtschaftsfreiheit. Schweizer Monatshefte 31(6): 333–338
Klinckowstroem W (2000) Walter Eucken: Eine biographische Skizze. In: Gerken L (ed) Walter Eucken und sein Werk. Mohr Siebeck, Tübingen, pp 53–115
Köhler EA, Kolev S (2013) The conjoint quest for a liberal positive program: "Old Chicago", Freiburg, and Hayek. In: Peart SJ, Levy DM (eds) F. A. Hayek and the modern economy: economic organization and activity. Palgrave Macmillan, New York, pp 211–228
Kolev S (2015) Ordoliberalism and the Austrian school. In: Boettke PJ, Coyne CJ (eds) The Oxford handbook of Austrian economics. Oxford University Press, Oxford, pp 419–444
Kolev S, Goldschmidt N, Hesse J-O (2014) Walter Eucken's role in the early history of the Mont Pèlerin Society, Discussion paper 14/02. Walter Eucken Institut, Freiburg
Köster R (2011) Die Wissenschaft der Außenseiter. Die Krise der Nationalökonomie in der Weimarer Republik. Vandenhoeck & Ruprecht, Göttingen
Nicholls AJ (1994) Freedom with responsibility. The social market economy in Germany, 1918–1963. Clarendon, Oxford
Oliver HM (1960) German neoliberalism. Q J Econ 74(1):117–149
Sally R (1998) Classical liberalism and international economic order: studies in theory and intellectual history. Routledge, London
Van Horn R (2009) Reinventing monopoly and the role of corporations: the roots of Chicago law and economics. In: Mirowski P, Plehwe D (eds) The road from Mont Pèlerin: the making of the neoliberal thought collective. Harvard University Press, Cambridge, MA, pp 204–237
Vanberg VJ (1988) "Ordnungstheorie" as constitutional economics. The German conception of a "social market economy". ORDO Jahrb Ordn Wirtsch Ges 39: 17–31
Vanberg VJ (2001) The Freiburg school of law and economics: predecessor of constitutional economics. In: Vanberg VJ The constitution of markets. Essays in political economy. Routledge, London, pp 37–51
White LH (2012) The clash of economic ideas. The great policy debates and experiments of the last hundred years. Cambridge University Press, Cambridge

European Community Law

Klaus Heine[1] and Anna Sting[2]
[1]Rotterdam Institute of Law and Economics, Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands
[2]International and European Union Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Abstract

European Community law has evolved over the last 60 years as a process of European economic integration. European law has developed alongside the stages of economic integration in a reciprocal process, with euro-area Member States being close to total economic integration. However, economic integration has not been unhalted – political and economic realities have forced the EU to adapt to current developments and negotiate treaty amendments but sometimes also to act outside the Treaty framework. Other influences, such as interest groups, have also shaped the dynamics of interest groups.

Synonyms

European law; European Union law

Definition

European community law consists of the law that is found in the Treaty on European Union (TEU) and the Treaty on the Functioning of the European Union (TFEU). In addition European community law comprises regulations, directives, or decisions, as well as the soft-law form of recommendations and opinions.

The Legal Framework of the European Union

Introduction: The History of Integration and Main Legal Texts

The European Union (EU) as we know it today is a result of a process of European integration that

has lasted for the most part of the last seven decades. While this might seem like a long time, it is only a blink of an eye in the overall history of the European continent.

The European Union has evolved from what was earlier called the European Community, which in turn was based on three communities founded in 1951 (the European Coal and Steel Community (ECSC)) and more importantly in 1957 (the European Economic Community (EEC) and the European Atomic Energy Community (EURATOM)) by means of the treaties of Paris and Rome, respectively (Lenaerts and van Nuffel 2011). While the Treaty on the Coal and Steel Community was concluded for 50 years only, the Communities as of 1957 were entered into for an unlimited period (art. 356 TFEU). Six Member States, Belgium, France, Germany, Italy, Luxemburg, and the Netherlands, committed themselves to the process of economic integration and gave away sovereign rights to a supranational order. They agreed to have some policy areas governed by four principal institutions: The Council as a representation of Member States, the Commission (or at that time the High Authority), the Assembly (later to be the European Parliament), and the Court. The Council and the Commission were triplicated for all three communities but joint for the purposes of efficiency and convenience by the Merger Treaty in 1965 (Craig and De Burca 2011; Hartley 2014; Lenaerts and van Nuffel 2011).

While the ECSC Treaty was very specific as regards the policy to be pursued by the institutions, the EEC treaty left much more leeway to the institutions to fill in their objectives. In addition, the mainly economic aims of the Communities were complemented by ideas of a political, social, and cultural union. Probably the greatest step for the Communities was the establishment of the European Union by the Maastricht Treaty in 1992, which also paved the way for the Economic and Monetary Union (Hartley 2014). This treaty was founded on the communities and introduced two more policy domains to be dealt with by cooperation: a common foreign and security policy and cooperation in the field of justice and home affairs (Craig and de Burca 2011).

However, the process of European integration did not only concern the deepening of integration but also the widening. Membership became open for other States in Europe: Even though the UK was initially reluctant to join the EU, it ended up applying for membership, first blocked by Charles de Gaulle but later realized in a group accession including Ireland, Denmark, and Norway in 1972. In the years 1979 and 1981, Greece, Spain, and Portugal joined, before some of the EFTA countries, Sweden, Finland, and Austria, acceded to the Union in 1994. The biggest accession process, however, concerned a group of former communist countries in the East of Europe (Lithuania, Latvia, Estonia, Poland, the Czech Republic, Slovakia, Hungary, Slovenia) and two Mediterranean islands, Cyprus and Malta (Nugent 2010). More recently, Bulgaria and Romania have joined and the currently last Member to the EU is Croatia, which acceded in July 2013 (Hartley 2014).

The Treaty of Lisbon, a result of the failed constitutional treaty for the European Union, replaced the European Community by the European Union as the sole entity and led to a deepening of the EU competences and codified institutional developments in the Union (Craig and de Burca 2011).

European Union Law as a Body of Law

Status of European Union Law and Main Legal Texts

The law of the European Union can roughly be divided into primary and secondary legislation. Primary law is the law that is found in the treaties, the Treaty on European Union (TEU) and the Treaty on the Functioning of the European Union (TFEU). The TFEU superseded the Treaty on the European Community as a result of the Treaty of Lisbon (Fairhurst 2012).

Secondary law is EU law that is adopted on the basis of primary law, the so-called legal bases, and can take the shape of regulations, directives, or decisions, as well as the soft-law form of recommendations and opinions (art.288 TFEU). Regulations have general application and are binding in their entirety, while directives are binding as to the result to be achieved (and need to be implemented by national authorities); however, Member States

are free to choose form and methods. Decisions are binding in their entirety on those to whom they are addressed. If there is no legal basis in the treaty, the EU is not allowed to adopt any legislation; this is called the principle of conferral as laid down in art.4 (1) and art.5 (1) TEU (Craig and de Burca 2011). Those areas in which the Union is allowed to legislate are divided in three categories:

Exclusive competence areas are those on which only the Union is allowed to legislate, the customs union, competition rules, monetary policy for euro-area Member States, conservation of marine resources under the common fisheries policy, and the common commercial policy (art.3(1) TFEU). This competence also applies in external context for the negotiation of international treaties, which the Commission is allowed to sign on behalf of the EU for these policy areas (art.3(2)TFEU).

Shared competence areas are those in which both the Union and its Member States are allowed to take legislative action and include the social policy; economic, social, and territorial cohesion; and transport (art.4(2) TFEU includes a full list). However, the level on which decisions are actually taken is governed by two important principles of the EU: subsidiarity and proportionality (Lenaerts and van Nuffel 2011). Both principles were put in the Treaty of Maastricht (TEU) and describe that:

> Under the principle of subsidiarity, in areas which do not fall under the exclusive competence of the Union, the Union shall act only if and in so far as the objectives of the proposed action cannot be sufficiently achieved by the Member States, either at central level or at regional and local level, but can rather, by reason of the scale or effects of the proposed action, be better achieved at Union level (art.5(3)TEU)
>
> and
>
> under the principle of proportionality, the content and form of Union action shall not exceed what is necessary to achieve the objectives of the Treaties. (art.5(4)TEU)

There is one competence that is the odd one out – economic coordination. While monetary policy is an exclusive competence of the EU and is conducted by the independent European Central Bank, Member States are to "coordinate their economic policies within the Union" (art.5(1) TFEU). This discrepancy in competence division is arguably one of the reasons for the euro crisis (de Grauwe 2012).

Not only can one categorize European Law on the basis of competences, another divide that is often made is the distinction between Internal Market law and institutional/constitutional law (see below). The notion Internal Market describes a Common Market between the members of the EU. While economic integration was initially the main purpose of the Union, the institutional setting has only evolved alongside the process of integration.

The Institutions

After the Treaty of Lisbon, the European Union is officially governed by seven institutions: the European Parliament, the European Council, the Council, the Commission, the Court of Justice of the European Union, the European Central Bank, and the Court of Auditors (art.13 TEU). These institutions have developed over time, although most of them were part of the Union framework even before they were officially mentioned as institution in the treaty. However, some institutional developments cannot go unmentioned (Hartley 2014).

The European Parliament used to be called the Assembly until 1962 (EP Resolution of 30 March 1962, JO1962) and hardly had any powers at all in decision-making – it was not even recognized as an institution. That changed in the case European Parliament v. Council Case 302/87 (1988) ECR 5615 concerning the Parliament's legal standing in front of the Court, which, before the judgment, was not "privileged," i.e. directly admissible with any action before the Court, a prerogative reserved for institutions (Hartley 2014). The European Parliament then started to play an ever greater role in the EU, culminating in the first direct elections for the European Parliament in 1979, and together with the national parliaments, it is now seen as the democratic underpinning of the European Union. It is comprised of currently 751 members representing the European people

by degressively proportionate representation, with a slight overrepresentation of small Member States.

More recently, the introduction of the so-called ordinary legislative procedure by the Treaty of Lisbon basically made the European Parliament co-legislator with the Council in most areas of European law (art. 294 TFEU). The Council, as the main legislator, is comprised of ministerial delegates who are authorized to commit their governments (art. 16(2) TEU); the Council does not have a fixed constellation, but the ministerial delegates will change depending on the topic to be discussed – finance ministers will legislate on financial matters, ministers of agriculture will legislate on agriculture, etc. The presidency of the Council rotates on a half-year basis among Member States; however the situation is different when it comes to a, one could call, special configuration of the Council, the European Council. The European Council is comprised of the Member States' heads of state or government and has, since the Treaty of Lisbon, a permanent president who is elected for two and half years. The European Council is the highest political organ of the European Union and provides it with the necessary impetus for its development and defines general political directions and priorities (art. 15 TEU; Craig and de Burca 2011). Especially in the euro crisis, the European Council has played a major role with regard to the measures taken to tackle the crisis, since the situation required a consensus on highest political level.

The Commission is an independent body representing the Union interest and is led by 28 Commissioners, one from each Member State. Commissioners are proposed by the European Council, while the European Parliament has to consent to the appointment of the president of the Commission and his or her Commissioners by the European Council (art. 17(5) TEU). The Commission is the only institution to have legislative initiative; it formulates proposals for new EU policies, mediates between Member States, and oversees the execution of Union policies. It could be regarded as the EU executive (Craig and de Burca 2011).

The European Central Bank (Hartley 2014, p. 31f) only became an official institution once the Treaty of Lisbon entered into force, even though it existed in the same shape before – since the inception of the Economic and Monetary Union (EMU). Its independence in conducting monetary policy for the Member States whose currency is the euro is enshrined in the treaty (art. 283(3) TFEU) and is only limited by its mandate being that of securing price stability (art. 127(1)TFEU).

The Court of Auditors consists of one national per Member State from within a pool of national auditors. Their task is to examine the accounts of revenue and expenditure of the European Union and to provide the Union legislator with reports on the regularity and legality of such (Hartley 2014, p. 31). The Court of Auditors is strictly a monitoring institution and is not to be confused with the Court of Justice of the European Union (CJEU). The CJEU has, in the context of the reforms of the Lisbon Treaty, undergone a major restructuring and is now divided in the General Court (a court of first instance and for the less difficult cases), the European Court of Justice (for seminal cases, it always sits in full court), and the Civil Service Tribunal (for cases involving the staff of the EU). The Court of Justice has played a major role in the process of European integration (Craig and de Burca 2011), which will be discussed below.

In general, however, the division of tasks among the institutions in the European Union is very well thought through and quite balanced. The Commission, together with national authorities represented in the European Council, acts as an executive; the Council and the Parliament now share the legislative competence, while the CJEU is the obvious judiciary. The European Central Bank is probably the only central bank whose independence is enshrined in a document of arguably constitutional value (the treaty). The democratic legitimacy of the European Union is based on the European Parliament and the indirect control that national parliaments have on their governments when they take decisions in the Council, as well as the principle of subsidiarity which is controlled by national parliaments (Lenaerts and van Nuffel 2011).

While the establishment of a democratic Union is one of the proclaimed aims of the EU, one has to be aware of a debate that has mainly started with the Treaty on European Union in Maastricht and that deals with the question of a democratic deficit in the European Union. Two strands of literature claim the existence of such a deficit. First, institutionally speaking, the European Parliament is argued to be unrepresentative of the European people since voter turnout is relatively low and the ordinary legislative procedure is not applicable in important policy areas such as the Common Foreign and Security Policy and the monetary policy. In addition, the involvement of national parliaments is often seen as too little (Follesdal and Hix 2005). Second, a sociopsychological democratic deficit is often attested for the Union. This refers to the lack of a common European people (demos), which in turn leads to a conceptual lack of a European democracy. Opponents on the other hand argue that there is no democratic deficit, because the Union is mainly a technocratic order and is also designed as such (Majone 1998). While this debate is worthwhile exploring, it would go beyond the scope of this entry to discuss it in detail.

European Union Law as a Legal Order

The treaties are essentially treaties that have been entered into under the rules of public international law. However, the European Union legal order has evolved as a legal order standing on its own. In this, the European Court of Justice has played a major role in the early years of integration. Two seminal cases defined the status of European Union law. First, in the case *Van Gend en Loos v Nederlandse Administratie der Belastingen* (1963) Case 26/62, which concerned a reclassification of a chemical in another customs category by the Benelux countries, the Court stated that the EEC Treaty was a legal order on its own, capable of creating legal rights which can be enforced directly by natural or legal persons before the courts of the Community's Member States, if the respective provision was "clear, precise, and unconditional." This principle is now called the principle of direct effect and is probably one of the most important principles of Union law and has been further developed (Craig and de Burca 2011).

Another seminal case that thrived European integration is the case *Flaminio Costa v ENEL* [*1964*] *ECR 585* (*6*/*64*) which regarded an alleged incompatibility of an Italian domestic law with the treaty. Here, the Court stated that "the law stemming from the treaty, an independent source of law, could not (. . .) be overridden by domestic legal provisions, (. . .), without being deprived of its character as community law and without the legal basis of the community itself being called into question." It thereby effectively established supremacy of the Union law over national law, which even applies in the case of national constitutions (*Internationale Handelsgesellschaft und Vorratsstelle für Getreide und Futtermittel* Case 11-70, *ECR 1970 1125*; Lenaerts and van Nuffel 2011).

The Dynamics of the European Union Law

The legal order of the EU has not only changed and developed because of the proactive Court of Justice. Rather, the EU has always reacted to social, economic, and political realities and has changed accordingly. Unfortunately, this has not always been to the benefit of a more coherent EU law. In order to change the treaties and provide the Union with more competences, all Member States have to be in agreement (art. 47 TEU) and ratification in all Member States is required. Increasingly, the Member States find it hard to reach consensus among, currently, 28. Some Member States are keen to drive the integration process forward towards a political and social union, while others are more reluctant. In some policy areas, therefore, some Member States have negotiated an opt out (such as the UK and Denmark regarding the monetary union) or those Member States that wanted to further integrate opted for different (intergovernmental) solutions (Craig and de Burca 2011).

The Schengen Treaty is one example of such further cooperation which started out as an international treaty outside the EU framework. Some Member States committed themselves to abolish border controls among themselves in order to facilitate the free movement of persons. Later,

however, the Schengen Treaty was incorporated into the EU treaty framework and is now part of the so-called acquis communautaire, the Union law that needs to be accepted and implemented by all new Member States. Only those Member States that did not ratify the Schengen Treaty before it was part of the Union framework (the UK and Ireland) are not obliged to comply with Schengen (Hartley 2014).

Similar developments have taken place with regard to the tackling of the economic and debt crisis. While EU measures enhancing the economic agreements under the so-called Stability and Growth Pact (SGP) have been adopted and implemented in the form of the six pack (six regulations and one directive) and the two pack (two regulations), agreement on the establishment of a permanent financial stability mechanism could not be reached. In addition, mechanisms for immediate financial assistance for Member States in case of an economic crisis were not foreseen in the treaty. The result was the establishment of somewhat hybrid institutions such as the European Financial Stability Facility, a *societe anonieme* under Luxembourg law backed up by euro-area Member States lending money to other euro-area Member States in financial trouble, the granting of bilateral loans to Greece, and the European Financial Stability Mechanism, all of which are now replaced by the permanent European Stability Mechanism, the ESM (de Grauwe 2012). The ESM as well is founded in a treaty outside the EU framework ratified by 25 Member States (apart from the Czech Republic, the UK, and Croatia), although a treaty change to art.136 TFEU has been made allowing the euro-area Member States to establish among themselves a European Stability Mechanism. Another intergovernmental treaty, the Treaty on Stability, Coordination, and Governance (TSCG), has been ratified by the same countries, in order to improve fiscal discipline and economic coordination among the signatories. Both treaties are designed to include the EU institutions such as the Commission and the Court in their working, and the EU aims to incorporate these treaties in the EU framework midterm. However, it is striking that intergovernmental decision-making in the context of treaties or the European Council has gained momentum in recent years. While it used to be the aim to bring more and more policy areas under the supranational pillar of decision-making (the ordinary legislative procedure), particularly the economic and monetary union is increasingly governed intergovernmentally. The European Council might be evolving from an institution that gives political impetus to an institution with a more legislative role. While this might arguably sidestep the European Parliament, it is also an example for the dynamics of EU institutional law. The mode of decision-making seems to increasingly depend on the sensitivity of the policy area concerned: Internal Market law on the other hand is still a prime example for supranational decision-making in the EU.

The Co-development of European Community Law and Economic Integration

The Economic Rationale for European Community Law and Integration

The body of European Community law implies a great promise for the European Member States, companies, and citizens. Tearing down barriers to trade and closer institutional integration can reap considerable welfare gains.

Even though the beneficial effects of free trade are known since the works of Adam Smith (1776) and David Ricardo (1817), it was the Cecchini-report from 1988 (Cecchini Report 1988) that estimated an overall cost reduction of 200 billion ECU for the Member States, if all trade barriers (tariff and non-tariff) would fall. However, the report provided no detailed analysis about the growth effects and distributional consequences for and between regions (e.g., center and periphery) (Baldwin 1989).

Because some Member States win and some lose with regard to their absolute welfare position, when trade barriers are removed (while overall welfare increases), it did not come as a surprise that in the aftermath of the Cecchini-report, the potentially losing Member States resisted a simple

abolishment of trade barriers (for a theoretical analysis, see Pierson 1996).

The insight that the distributional consequences of abolishing trade barriers which must be taken into account leads to three basic propositions that drive the economic analysis with regard to European Community law.

1. Economic integration is a process in time in which the Member States go through *stages of integration*, thereby deepening integration.
2. European Community law is mirroring the stages of integration, thereby aiming towards a European *constitution*.
3. At the micro-level the integration process is driven by *interest groups* and stakeholders that impact on the content of European Community law.

Stages of Integration

It is common to differentiate five generic stages of economic integration (Balassa 1961; Baldwin and Wyplosz 2012). Economic integration usually starts at the first or second stage and progresses then over time towards stages of higher integration.

1. The first stage is mainly associated with a bilateral or multilateral free trade agreement (FTA) between countries. While the abandoning of tariffs between the signatories of the FTA reaps some welfare gains, running an FTA is not trivial and not easily administered. That is because of the FTA countries' freedom to determine the tariff with third countries on their own. For importers and FTA countries alike, this creates opportunities for arbitrage (Tarr 2009). As a consequence rules of origin have to be introduced and maintained. One may conceive these rules of origin as a very nascent step into the direction of a common understanding over the rules of the game between FTA countries. One may regard this as a first step towards a premature constitution.
2. The second stage is the customs union (CU). The signatories of a CU abandon tariffs against each other, but they also agree on a common tariff against third countries. This overcomes the problem of maintaining and controlling complex rules of origins as it is necessary under an FTA. From an economic perspective, the Treaty of Rome (1957) constituted a CU.
3. The third stage is the establishment of a common market (CM). The Treaty of Rome implied already the antecedents of a CM, although it took time till the 1980s to establish and enforce the features of a CM. Central to a CM is that non-tariff barriers to trade (national regulations) which impede trade become removed and that the free movement of people/companies, services, and capital becomes feasible. Two features of a CM need special attention: First, non-tariff barriers for goods, people/companies, services, and capital (four freedoms) can either be overcome by harmonizing laws and regulations or by mutual recognition of laws and regulations. Both ways play out in the European economic integration process. Second, to safeguard the four freedoms, a CM has to establish a common competition policy, preventing the Member States to give unfair advantages to their industries (esp. state aid control). Both features have in common that they need policy coordination between Member States, implying that there is a basic agreement about the vertical delineation of the competencies between the central level (EU level) and the signatories (Member States). In parallel power has to be vested to institutions and administrations at the central level, in order to execute the common policies. This process of institutionalization becomes apparent, for example, in the increasing role that the Court of Justice of the European Union plays.
4. The fourth stage is the economic union (EUN), which is a CM where a number of key policy areas are harmonized (esp. coordination of monetary and fiscal policies, labor market, regional development, infrastructure, and industrial policy). While at the first three stages of economic integration the removal of obstacles to free trade are center stage (negative integration), at the fourth stage the particular institutional design of the economically integrated area becomes important. This process of

institutionalization goes hand in hand with the strengthening of administrative bodies managing for the particular design of integration (positive integration). However, the implementation and enforcement of policies remain with the associated countries. The establishment of the European Single Market (Maastricht Treaty) is a good example of a EUN.

5. The fifth stage is total economic integration (TEI). At this stage, a vast number of key policies become harmonized, inter alia those as social policy and monetary and fiscal policy. A TEI exploits all possible economic advantages from free trade and factor mobility. The momentum of positive integration becomes amplified by shifting even more power to the central administration. The policies from the central level become binding for the associated countries, which role is only to execute the prescribed policies from the central level. The EU has yet not reached the stage of TEI, but the members of the euro area are close to it.

Towards a European Constitution

European Community law has yet not been consolidated in a European constitution. The "Treaty establishing a Constitution for Europe" from 2004 was not ratified by all Member States; as a consequence in 2005, the project of a European constitution was given up. Instead in 2009 the Treaty of Lisbon implemented parts of the constitutional project (e.g., qualified majority voting) through the established European treaties.

Insofar, in a strict legal sense, European Community law has yet not reached the status of a constitution. However, European Community law has undoubtedly evolved over time towards a tighter set of rules and regulations (institutional integration). From a law and economics perspective, the crucial question is whether the institutional integration of the EU corresponds indeed with more transborder transactions and hence more gains from trade. In other words, does the process of European constitutionalization lead to welfare gains?

This question can only be answered empirically and it bears the question of how to measure institutional integration. However, Mongelli et al. (2005) can show that a higher degree of institutional integration goes hand in hand with more economic integration (trade). Thereby it seems that causation runs both ways: More institutional integration triggers more economic integration, but more economic integration also leads to more institutional integration. From that observation follows that there is a coevolution between European Community law and economic integration. This implies that the European constitutionalization process is driven by policy choices for the background of a deepening economic integration, while at the same time the deepening of economic integration depends on the policy choices that were made.

The Role of Interest Groups

European Community law is the product of actors on the national and the EU level which try to shape the governing rules and the law-making process into a for them favorable direction. Usually those actors are labeled as interest groups. Even though the term "interest groups" must not have a negative connotation but can refer to quite a lot of meanings in the process of European integration (Eising 2008), from an economics perspective it refers mainly to well-organized groups that try to get an economic advantage that is not the result of their economic performance but of their successful attempts to get favors via the political process (rent seeking) (Mueller 2003; with special reference to federations, see Weingast 1995). For example, the persistent high levels of subsidies that go into the common agricultural policy (CAP) can only be explained by the influence that the interest group of farmers has on European politics (Baldwin and Wyplosz 2012).

The role of interest groups in the process of European integration in general and the impact of interest groups on European Community law in particular can be studied from a great number of perspectives (for an overview, see Eising 2008). From a law and economics perspective, it is important to understand the making of European Community law as the result of a coevolutionary process of diverse interest groups, which influence laws and regulations, and the body of European law, which creates the forum in which

interest groups and stakeholders can play out. Thereby the European Community law becomes over time more differentiated and adapted to the challenges it faces in an ever more economically integrated area, but at the same time the increasing complexity of European Community law creates gaps and loopholes that can be targeted by interest groups. This implies on the one hand that interest groups have to evolve to organizations that are able to target those loopholes and gaps. On the other hand, it implies that Community law reacts to the activities of interest groups with a further differentiation of law and regulations.

While the relevance of interest groups for the development of European institutions and Community law cannot be doubted, the operation of interest groups in the EU is yet not fully understood. For example, it is yet not clear whether there is a Europeanization of interest groups (making the level of Member States less important) or whether successful interest groups on the European level need a strong anchoring in the Member States. Moreover, one may ask whether the specific politico-legal system of a Member State predetermines the effectiveness of interest groups on the European level (Eising 2008).

References

Balassa B (1961) The theory of economic integration. Irwin, Homewood

Baldwin R (1989) The growth effects of 1992. Econ Policy 4:247–282

Baldwin R, Wyplosz C (2012) The economics of European integration, 4th edn. McGraw-Hill, London

Cecchini-report (1988) The European challenge. Gower, Aldershot

Craig P, de Burca G (2011) EU law- text, cases and materials, 5th edn, Oxford University Press, Oxford

De Grauwe P (2012) The governance of a fragile eurozone. Aust Econ Rev 45(3):255–268

Eising R (2008) Interest groups in EU policy making. Living Rev Eur Gov 3:4–32

European parliament resolution of 30 March 1962, JO1962

Fairhurst J (2012) Law of the European union law, 8th edn. Pearson, Harlow

Follesdal A, Hix S (2005) Why there is a democratic deficit in the EU. European Governance papers (EUROGOV) C-05-02

Hartley T (2014) The foundations of European union law, 8th edn, Oxford University Press

Lenaerts K, Van Nuffel K (2011) European union law, 3rd edn. Sweet & Maxwell, London

Majone G (1998) Europe's 'democratic deficit': the question of standards. Eur Law J 4(1):5–28

Mongelli FP, Dorrucci E, Agur I (2005) What does European institutional integration tell us about trade integration? ECB occasional paper series 40

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

Nugent N (2010) The government and politics of the European union, 7th edn. Palgrave MacMillan, Basingstoke

Pierson P (1996) The path to European integration: a historical institutionalist analysis. Comp Polit Stud 29:123–163

Ricardo D (1817) On the principles of political economy and taxation. John Murray, London

Smith A (1776) An inquiry into the nature and causes of the wealth of Nations. Methuen, London

Tarr RD (2009) Customs union. In: Reinert KA, Rajan RS (eds) The Princeton encyclopedia of the world economy, vol 1. Princeton University Press, Princeton, pp 256–259

Weingast B (1995) The economic role of political institutions: market-preserving federalism and economic development. J Law Econ Organ 20:1–31

Further Reading

Chalmers D, Davies G, Monti G (2010) European union law, 2nd edn. Cambridge University Press, Cambridge

Moravcsik A (2002) In defence of the democratic deficit: reassessing legitimacy in the European Union. J Common Mark Stud 40(4):603–624

European Court of Human Rights

Federica Gerbaldo

IEL- Institutions, Economics and Law, Collegio Carlo Alberto, University of Torino, Turin, Italy

Abstract

The European Court of Human Rights is a Supra-national Court established in 1959 with the European Convention on Human Rights. It stands as a monitoring mechanism to ensure the observance of the commitments to ensure fundamental human rights undertaken signing the Convention.

This entry focuses on the functioning mechanism of the Court either from a prescriptive and an empirical point of view. The first aspect

concerns organization and procedural issues, with special regard to the sanctioning mechanism, whereas the second aspect involves issues of effectiveness of the Court.

Definition

The European Court of Human Rights (ECtHR) is a supranational court founded by the European Convention of Human Rights (ECHR) in 1959 and based in Strasbourg (France). It carries out a supervisory function monitoring that the 47 member states of the Council of Europe (COE) that have ratified the Convention comply with its substantive provisions. The court fulfills its task scrutinizing claims alleging that the defendant state breached its commitment violating one or more ECHR provisions.

History

Founded in 1949, the Council of Europe (COE) became the very first political organization in Europe, although its approach was different from the European institutions. In the era of reconstruction after World War II, the original member states chose the path of cooperation and signed the founding treaty of COE as a reaction to the atrocities of the war and the growth of East-West tension. States parties made efforts in strengthening cooperation, signing international treaties, publishing peer reviews, exercising pressure, and promoting training and good practice; reinforcing citizenship and democratic governance was the ultimate goal.

The ambitious project of *achieving a greater unity between its members for the purpose of safeguarding and realizing economic and social progress* (article 1 Statute of council of Europe, 1949. Add to documents refrences Council of Europe, Statute of the council of Europe, 05 May 1949) aimed to guarantee lasting peace and prosperity on a continental scale. In this sense, the most successful achievement is the ECHR signed in 1950 and entered into force in 1953 and the creation of a monitoring mechanism with the establishment of ECtHR. As a response to the inertia of the United Nations and its inconclusive attempts to transmit the principles proclaimed in the 1948 Universal Declaration of Human Rights into an internationally binding bill of rights, some of the Western European countries egged on the COE to proceed on its own. Thanks to the obduracy of those countries that share *a common heritage of political tradition*, *ideals*, *freedom and the rule of law* (European Convention of Human Rights, preamble, 4th November 1950, 213 UNTS, 221), the COE took the first steps for the collective enforcement of some of the human rights stated in the Universal Declaration.

Even though ECHR scope is more specific, the Convention arranges one of the strongest regional mechanisms for protection of fundamental individual human rights and stands as a model for other regional systems (Buergental 2006). Signing the Convention, high contracting parties undertake to grant to every individual within their jurisdiction civil liberties typical of an effective political democracy. Among others, the Convention grants life, fair hearings, private and family life, freedom of expression, freedom of thought, conscience and religion, and property rights and prohibits torture or inhuman treatment, slavery and forced labor, arbitrary and unlawful detention, and discrimination. Besides the declaration of a list of human rights that contracting states commit to recognize and grant (see Harris et al. 2014 for an analysis of the rights granted), the Convention made protection effective establishing a monitoring mechanism to *ensure the observance of the engagement undertaking by the high contracting parties* and providing remedies for victims of violations.

After the entry into force, member states implemented the Convention with Protocols oriented to adjust structure and functioning mechanism to the constant increasing workload of the court (Egli 2007, p. 7). According to the original design, the court envisaged a double-step mechanism: a commission performed monitoring tasks analyzing claims' conformity to admissibility criteria and attempting friendly solutions, whereas a court performed adjudicatory tasks. The accession to ECHR of new member states after the fall of the Berlin Wall dramatically increased the

number of claims: Protocol 11, entered into force on 1 November 1998, was imagined to deal with the huge number of pending applications. The articulated mechanism moved from a mix of monitoring and adjudicatory tasks to an entire judicatory system, based on a single court working on a permanent basis: the elimination of the commission screening function simplified and made the procedure more concise. Furthermore, it allowed individual applicants to present claims directly to the court.

Despite its purpose, Protocol 11 failed to deal with the ECtHR workload, and a number of applications incessantly increased during years. In 1999, right after the introduction of Protocol 11, the number of pending applications was 12,600; 10 years later, pending applications amounted to 119,300 (ECtHR statistics 2013, p. 7). Member states issued Protocol 14, which entered into force on 1 June 2010 and provided further reforms aiming to grant long-term efficiency and strengthening applications' process. The protocol reduced the decision body formation for the simplest claims, conditioned the admissibility of claims to further admissibility criteria requiring "a significant disadvantage" for the applicants, and prolonged the judges' term of office from 6 years with option of reelection to nonrenewable 9 years term providing for expiration of the office when they reach 70 years of age.

Currently, members have adopted Protocol 15 introducing references to the principle of subsidiarity and the doctrine of the margin of appreciation and reducing from 6 to 4 months after the date of the final domestic decision the time limit within which the claims should be filed and Protocol 16 allowing the court to give advisory opinions on questions of principle related to interpretation and application of rights, upon request of the highest domestic courts. Both protocols will enter into force only once contracting states will sign and ratify them.

Organization

Issuing durable legal rules per se does not make a regional human rights system effective, if there is no assignment to third parties of the task of interpreting and applying rules. ECtHR carries out exactly this role and makes the international regional system established by ECHR capable of enforcement. The effects of the incorporation of ECHR together with the functioning of the court as a constitutional court and the effects of ECtHR judgements allow denial of those theories that support a qualitative difference between international regimes and states; thus doubting some form of constitutionalism beyond states is feasible (Rosenfeld 2010). Recent theories argue that the ECHR framework evolves into a transnational constitutional regime (Stone Sweet 2012).

The ECtHR has its headquarters in Strasbourg and is composed by elected judges, chosen either among judges eligible for high judicial office or jurists of esteemed competencies. The number of judges is the same as the number of the high contracting parties (47 at present). The Parliamentary Assembly elects judges with majority of votes, among a shortlist of three candidates presented by each contracting party.

Judges must have high morality: although representing one state, they hear the claims in their individual capacity. Thus, they must avoid any activity which endangers their impartiality and independence.

Depending on the complexity of the case, the ECtHR scrutinizes the claim sitting in one of its four judicial formations. When inadmissibility of the claim clearly appears without any further investigation, the single judge has the power to strike the complaint out of the list of the cases or declares its inadmissibility with final decision. When this is not the case, the single judge forwards the claim either to a Committee or to a Chamber. The Committee unanimously decides either to declare inadmissible the claim or strike it out when no further investigation is required or to declare it admissible and simultaneously issue the judgement on the merit when the question is the object of well-established jurisprudence. Alternatively, the Chamber rules on the merit of the claim, together with a decision on admissibility or separately. In the exceptional situation in which the case concerns a question seriously affecting ECHR interpretation or which might

lead to a decision contrasting with a previous judgements, the Chamber relinquishes jurisdiction in favor of the Grand Chamber of 17 judges. Moreover, one or both parties might require referral to the Grand Chamber within a period of 3 months from Chamber judgement's delivery.

Procedural Issue

ECtHR has jurisdiction to hear claims filed by individuals or state and alleging that one of the high contracting parties violated the ECHR. Jurisdiction includes all the matters related to interpretation and application of the ECHR; however, the court shall not take case on its own motion. Since the introductory phase, the procedure stands out for its accessibility and effectiveness. Both states and individuals, either person, group of people, or non-governmental organizations, have the right to present claims, thus granting widest access to justice.

A complaint undergoes two steps, with some faint adjustments concerning the analysis depending on the judiciary formation scrutinizing it. The first is the admissibility stage during which judges verify the observance of some formal requirements.

Claims must concern one of the rights covered by ECHR, not to be manifestly ill-founded and be direct against one or more states parties of the Convention. Following a general rule of International Customary Law, applicants shall exhaust all the available domestic remedies, up to the highest level of jurisdiction, and resort the ECtHR within 6 months following the date of the last judicial decision. Last, the applicants shall have suffered a significant disadvantage: the violation, although real from a mere legal point, shall overtake a minimum threshold to deserve international judges' attention. These requirements reflect different purposes: the former grants national authorities the opportunity to prevent, or at least to fix, violations, whereas the latter facilitates the activity of the court making it easier when dealing with unmeritorious claims and allowing devotion of more time to serious claims. Both rest on the same assumption: the subsidiarity of ECtHR with respect to national mechanism of safeguarding human rights. Thus, on one side, national courts have at first instance the chance to deal with questions regarding the compatibility of domestic law with the Convention. On the other side, the court stands as a supranational court of final instance deserving the analysis of substantial claims and denies the role of an alternative channel to which require further monetary damages.

Past the admissibility screening, judges scrutinize the merit of the claim and undertake further investigation, if needed, inviting parties to submit further evidence and written observations or fixing public hearings with the representatives of the parties. At the end of the investigation, judges deliberate on the question and issue a judgement, providing reasons for the decision, which becomes final if parties declare to renounce or do not request referral to the Grand Chamber within 3 months.

In principle, the final judgement is declaratory in nature (Marckx v. Belgium, appl. number 6833/74 1979 par. 58). Only in 2011, the court introduced an original mechanism aiming at optimizing time and resources allocation when dealing with cascade of repetitive claims, which does not necessarily end with a declaration that a violation has been committed. When applications underline the existence of a structural or systemic problem, the court might issue a pilot judgement after processing them as a matter of priority. The pilot judgement contains both the identification of the nature of the structural problem or the dysfunction and the suitable remedies the defendant state is required to adopt.

Final judgements are binding upon the parties of the cases. The declaratory character implies that the court renounces to impose on the breaching state any obligation to adopt specific measures necessary to ensure compliance, a part for some exceptional cases. The execution of the final judgement consists into the performance of two groups of obligations, depending on the measures required. The first one demands adoption of general measures and includes the obligation to execute the violated provisions and to prevent the occurrence of further violations. The second one consists into the obligation to put an end to the

violation and fix the negative effect, the execution of which required individual measures.

General measures aim to prevent future violations similar to the ones found by the court and most of the time require changes in legislation. However, when national authorities recognize direct effect to judgements, publication and dissemination of the decision is sometimes enough to induce national effective remedies.

Individual measures aim to put an end to the violation and to *restore as far as possible the situation existing before the breach* (Brumarescu v. Romania 1999, par. 19). Thus, following customary rules of international law, restitutio in integrum, aiming to take the injured back as far as possible in the same situation as the one he enjoyed before the violation, represents the preferred relief. When the restoration of the injured victim is otherwise impossible or only partial, the court might award monetary compensation.

The Committee of Ministers is the organ charged of supervising the execution of the judgements. The Committee organizes two meetings a year and gathers one representative for each contracting party (in principle the Minister of Foreign Affairs), each of whom is entitled to one vote. The monitoring function of the organ includes the supervision of the execution of friendly settlement and judgements (White and Ovey 2010, p. 53). The Committee invites the respondent state to explain the measure taken to avoid the consequence of violation and provide evidence of compensation payment: the Committee supervises regularly each case until it is satisfied with the adoption of general measures necessary to ensure compliance.

Just Satisfaction

Besides the general freedom in the execution of the judgements, the ECHR provides the court with the power to award compensation for injuries suffered, as a residual remedy for those damages that cannot otherwise be repaired. The core of the system lies in article 41 that provides for just satisfaction. The provision envisages that if a violation is found and *if the internal law of the contracting party concerned allows only partial reparation to be made*, *the court shall*, *if necessary*, *afford just satisfaction to the injured party.* The court might exercise the power to afford just satisfaction to victims in the same judgement if the question is ready for decision and claimants specifically require monetary compensation. Just satisfaction is a compensation for an actual harm: finding of evidence of violation does not automatically grant a positive award, being a matter of court discretion. The letter of the provision is clear when specifying that the award is not a direct and automatic consequence of the finding of violation. Judges determine to award just satisfaction upon a specific applicants' request supported by appropriate documentary evidence. The approval of the request is conditional on the inadequacy of national remedies to provide full reparation to victims. In addition to subsidiarity, the award shall be necessary and just.

Economic analysis of law acknowledges damages as an instrument having twofold function: restoring the injured for the harm suffered (compensation) and providing injurers with behavioral incentives fostering internalization of the externalities (deterrence). Due to their economic function, in general, damages shall provide full compensation for the victims' losses because only in this case the injurer internalizes the negative externalities produced (Posner 2003, p. 192). Nevertheless, compensation of damages within the international public field assumes a peculiar meaning considering that the court exercises only a weak form of review (Bernhardt 1994, p. 297), but aims to provide "constitutional justice," rather than "individual justice." On one side, when finding a violation of the Convention, judges' declaration of incompatibility with ECHR does not replace the questioned law with one of its own making, nor directly affect the validity of that law in the national legal system. On the other side, the actual function of the court is to ensure that administrative and judicial processes in member states effectively conform to pan-European convention standards, rather than seeking to provide every deserving applicant

with a remedy for the Convention violation (Greer 2003, p. 405).

Traditional economic theories argue that damages do not have to fully compensate harm, as long as they make taking due care the most attractive strategy (Visscher 2009, p. 155). However, having regard to the international human rights protection framework, under-compensation of the harm caused might not be sufficient in view of obtaining the ultimate goal of the court, namely, the adoption of a standard of protection appropriate to the ECHR. On the other side, the rules specifically clarify that the court does not aim to punish the breaching state for its responsibility in violating the Convention and invite to charge as inadequate any claims for *punitive*, *aggravated or exemplary* damages. Just satisfaction turns out to be the result of a balance between the interest of injured applicants to receive some sort of satisfaction and the public interest of defendant states, with an eye to the local economic circumstances of the state.

The award, if any, is exclusively in the form of a sum of money to be paid by the respondent state within a time limit that usually the court sets in 3 months. Judges might either award a global sum, either breakdown just satisfaction into three components that correspond to material losses, moral losses, and costs and expenses. Practical indications on the computation of material damages refer to the principles of International Customary Law codified by the United Nations Law Commission (chapter IV.E.1, article 28 and following). The basic principle rests upon the assumption that the injured applicant *should be placed*, *as far as possible*, *in the position in which he or she would have been had the violation found not taken place*. Material damages awards concern economic losses and would in principle reflect *the full calculated amount of damages* or a reasonable estimation based on observable facts.

Moral damages include a mixture of highly intertwined elements, many of which involved social and psychological aspects: it is a matter of civil law jurisdiction, rather than an instrument of public international law (Parish et al. 2011, p. 225). This concept recurs in context of tort that causes intangible, or otherwise difficult to quantify, injuries to a person and/or his/her rights, including pain and suffering, anguish and distress, and loss of opportunities. Assessment of moral damages is difficult because moral losses are not directly visible (Shavell 2004, p. 242; Aarlen 2013, p. 439). The court itself admits that *it is in the nature of nonpecuniary damages that it does not lend itself to precise calculation* (Practice Directions, section III, par. 13). As a matter of fact, assessment of damages grounds on equitable arguments and standards emerged by case law.

Empirical Studies Focusing on ECtHR

ECtHR counts the biggest caseload among international courts, with more than 17.000 judgements delivered since its creation, and makes most of the information about the cases publicly available. This makes the court as a suitable object of attention when investigating the more relevant issues related to effectiveness of international adjudication.

Voeten (2008) investigates international judicial behavior, focuses on impartiality of ECtHR judges, and considers the court as a committee of individual judges, rather than a single actor. Focusing on the minority opinions expressed by judges, he evaluates what determinants other than law, if any, systematically affect judges. Findings show that ECtHR judges' behavior is more similar to that of national review courts, rather than political institutions: there is no evidence that judges' domestic legal culture is a matter of concern, and the overall effects of career motivation are moderate (despite some evidence that judges show national bias in evaluating politically sensitive case). The study is overall positive toward the possibility of impartial review and suggests ECtHR judges are politically motivated actors having political preferences on to the best way to apply abstract human rights, but they do not use their power to settle geopolitical scores.

Helfer and Voeten (2014) take the ECtHR rulings on a specific issue, namely, lesbian, gay,

bisexual, and transgender rights, to investigate the effectiveness of international tribunals as agent of policy changes. The study aims to investigate whether the court affects the behavior of actors other than parties to a dispute or, contrary, the court simply reflects evolving social trends and adapts its jurisprudence to the evolution of national policies. The study contributes with an empirical assessment to the path of literature trying to provide analytical explanation of the effectiveness of international adjudication activities (Helfer 2014). In particular, it focuses on what has been classified as the *erga omnes* effectiveness and *norm development effectiveness*. Findings show that international law created by ECtHR jurisprudence systematically affects domestic judicial review and legal changes. ECtHR judgements increase the likelihood that all COE members, even those that are not found to violate the Convention, adopt friendly LGBT rights reforms, and the effect is much more evident where support for LGBT rights is relatively low. Based on statistical findings, the authors argue that the ECtHR court contributes to legitimize and justify political changes especially with respect to lazy countries. The court engages in majoritarian activism, rather than aggressive policy, and reverses previous decision only when at least a majority of COE members have already done so.

References

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

Bernhardt R (1994) Human rights and judicial review: the European Court of Human Rights. In: Beatty DM (ed) Human rights and judicial review: a comparative perspective. Martinus Nijhoff Publishers, Dordrecht, pp 297–319

Buergenthal T (2006) The evolving international human rights system. Am J Int Law 100(4):783–807

Egli P (2007) Protocol no. 14 to the European convention on human rights and fundamental freedoms: towards a more effective control mechanism? J Transl Law Policy 17(1):1–34

Greer S (2003) Constitutionalizing adjudication under the European Convention on Human Rights. Oxf J Leg Stud 23(3):405–433

Harris D, O'Boyle M, Bates E, Buckley C (2014) Law of the European convention on human rights, 3rd edn. Oxford University Press, Oxford

Helfer LR (2014) The effectiveness of international adjudicators. In: Romano C, Alter K, Shany Y (eds) The Oxford handbook of international adjudication. Oxford University Press, Oxford, pp 464–482

Helfer LR, Voeten E (2014) International courts as agents of legal change: evidence from LGBT rights in Europe. Int Organ 68(01):77–110

Parish MT, Nelson AK, Rosenberg CB (2011) Awarding moral damages to respondent states in investment arbitration. Berkeley J Int Law 29(1):225–245

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

Rosenfeld M (2010) The identity of the constitutional subject: selfhood, citizenship, culture and community. Routledge, London, pp 61–65

Shavell S (2004) Foundations of economic analysis of law. Harvard University Press, Harvard, p 242

Stone Sweet A (2012) The European convention on human rights and national constitutional reordering. Cardozo Law Rev 33(12):1859–1868

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

Voeten E (2008) The impartiality of international judges: evidence from the European Court of Human Rights. Am Polit Sci Rev 102(4):417–433

White RCA, Ovey C (2010) White and Ovey. The European convention on human rights, 5th edn. Oxford University Press, Oxford

Documents

Brumarescu v. Romania, appl. number 28342/95 (28 Oct 1999) 1999-VII; 35 EHRR 887. Available at http://hudoc.echr.coe.int/sites/eng/pages/search.aspx?i=001-58337

Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 Nov 1950, ETS 5. Available at http://www.echr.coe.int/Documents/Convention_ENG.pdf

Council of Europe, Rules of the court, 14 Apr 2014. Available at http://www.echr.coe.int/Documents/Rules_Court_ENG.pdf

European court of Human Rights, Analysis of statistics 2013, Jan 2014. Available at http://www.echr.coe.int/Documents/Stats_analysis_2012_ENG.pdf

Marckx v. Belgium, appl. number 6833/74 (13 June 1979) series A n. 31. Available at http://hudoc.echr.coe.int/sites/eng/pages/search.aspx?i=001-57534

Rules of the court, practice directions, just satisfaction claims. Available at http://www.echr.coe.int/Documents/PD_satisfaction_claims_ENG.pdf

United Nations Law Commission, Draft articles on responsibility of states for internationally wrongful acts, Nov 2001, Supplement Nr. 10 (A/56/10). Available at http://legal.un.org/ilc/texts/instruments/english/draft%20articles/9_6_2001.pdf

Further Reading

Bates E (2010) The evolution of the European convention on human rights. From its inception to the creation of a permanent court of human rights. Oxford University Press, Oxford

Brems E, Janneke G (2014) Shaping rights in the ECHR: the role of the European court of human rights in determining the scope of human rights. Cambridge University Press, Cambridge

Fenyves A, Karner E, Koziol H, Steiner E (eds) (2011) Tort law in the jurisprudence of the European Court of Human Rights. Walter DE Gruyter, Berlin

Greer S (2006) The European convention on human rights achievements, problems, and prospects. Cambridge University Press, Cambridge

Harmsen R (2007) The European Court of Human Rights as a "Constitutional Court": definitional debates and the dynamics of reform. In: Morison J, McEvoy K, Anthony G (eds) Judges, transition and human rights cultures. Oxford University Press, Oxford, pp 33–53

Keller H, Sweet AS (2008) A Europe of rights: the impact of ECHR on national legal system. Oxford University Press, Oxford

Senden H (2011) Interpretation of fundamental rights in a multilevel legal system: an analysis of the European Court of Human Rights and the Court of Justice of the European Union. Intersentia, Cambridge

Xenos D (2012) The positive obligations of the State under the European Convention of Human Rights. Routledge, London

European Identity

► European Nationality

European Integration

Aurelien Portuese
University of Westminster, London, UK

Abstract

European integration is a process. It is a process of institutionalization by which the old continent is continuously transformed either through incremental changes or through across-the-board reforms. European integration designates the changing current institutional framework of the European Union (hereafter EU). The European continent experiences other institutional integration than the EU's – mostly the European Convention on Human Rights – but the focus shall be put on the EU as this regional organization is the most significant and integrated one.

The history of the European integration has paved the way for an entire new legal order with a strong economic rationale to come to the fore (I). The emphasis of institution-building and market-building has nevertheless not been without difficulties in terms of construing a common political structure, democratically legitimate for the European national democracies (II).

Definition

European integration is an expression describing a process, by opposition to a state of affairs that would be mainly unchanged. Hence, the European integration defines an institutional process taking place within a regional organization by opposition to an institutional state of affairs created suddenly through different forms of state such as a federation, a regional state, a nation-state, etc.

The EU is not a state; it is an organization so much of its own that it is commonly described as being a sui generis organization. The EU should not be looked at or investigated as photography but rather as a movie in the making – hence the expression "European *integration*."

The history of the European integration has paved the way for an entire new legal order with a strong economic rationale to come to the fore (I). The emphasis of institution-building and market-building has nevertheless not been without difficulties in terms of construing a common political structure, democratically legitimate for the European national democracies (II).

The Law and Economics of European Integration

The history of the European integration portrays a strong economic underpinning of the common

destiny the European states have decided to embrace (a). Out of the economic system set up in the aftermath of the Second World War, an ongoing process of institutionalization has continuously transformed the European integration to become nowadays the most institutionalized region of the world. This process has spawned a new legal order – the European Union (EU) law – which is imbued with economic considerations (b). The European institutional framework and practice that result from this evolution is, from a law and economics perspective, driven by an economic efficiency rationale (c).

Economics of European Integration: Peace and Prosperity Through Reciprocal Stakes

On a continent repeatedly ravaged by wars and erstwhile enmities, and on a continent continuously disheartened by peace treaties violated, new peace-making solutions have appeared to become imperative during the postwar period. Diplomatic relations had not been so far satisfactory: only economic interdependence would make peace durable.

Because peace leads to prosperity, the argument of this Kantian perspective is circular: an economic integration paves the way for an ever-greater economic interdependence among European states allowing for peace and prosperity on the continent (Kant 1992).

The interlinkage between commercial knots and peace-building is well known since Montesquieu wrote: "The natural effect of commerce is to bring peace. Two nations that negotiate between themselves become reciprocally dependent, if one has an interest in buying and the other in selling. And all unions are based on mutual needs" (Montesquieu 1758, p. XX).

Therefore, the architects of the European integration were driven by fostering the self-interest of each state to make peace: the challenge was not so much about having a possible peace treaty to be signed but more about having war made impossible. Indeed, the economic interdependence created by free trade would render war impossible among rational and self-interested states.

The birth of European integration has consequently been triggered not by setting up a "superstate," a federation, or any other grand projects against which nation-states were reluctant to build. It has been driven by rational choice perspective and by the progressive deconstruction of the protectionist powers of the state.

Without coincidence, the first sectors of the economy to be integrated into a common pool were the coal and steel of European nations since coal and steel were the raw material for wars. The Schuman Declaration of 1950 in which Robert Schuman, French Minister of Foreign Affairs, called for a "step-by-step" process of integration leading to a European federation. This declaration heralded the creation in 1951 of the European Coal and Steel Community (ECSC), with a High Authority managing the production and the distribution of coal and steel in Europe. Consequently, with the ECSC, war has been rendered materially impossible (Pelkmans 2006).

In light of the success of the ECSC, the founding members – France, Germany, Italy, Belgium, Luxembourg, and the Netherlands – generalized this success to all sectors of the economy by creating a free trade area through the abolition of customs duties and the proclamation of the free movement of workers, capital, goods, and services. This was the creation of the European Economic Community (EEC) in Rome in 1957 (Duchene 1994).

The common market laid down by the EEC was constituted of a customs union made of free trade area (abolition of tariffs within Member States) and a common external tariff with respect to non-Member States. Also, the four economic freedoms (free movement of goods, services, workers, and capital) were proclaimed. Finally, few common policies were envisaged, such as the Common Agricultural Policy.

There are four stages of economic integration: a free trade area, a customs union, a common market, and finally an economic and monetary union. Before embracing a monetary union with the creation of the Euro in 1992 with the Maastricht Treaty, the EU has therefore jumped from its inception in 1957 into a very integrated regional union – a common market with a common external tariff and with common policies.

A 12-year transition period was granted in order for Member States to abolish their custom tariffs. In 1969, the common market was accomplished. The removal of barriers to trade and the large-scale competition allowed for over the entire continent would stimulate modernization of the economies by the lowering transaction costs, the enjoyment of economies of scale, and the enhancement economic efficiency. The belief in the net positive effects of a customs union was widely shared among economists.

The most influential theory at that time was Jacob Viner's theory of trade creation and trade diversion. He argued that economic integration leads to two simultaneous welfare effects: trade creation among Member States and trade diversion among Member States. Whether trade creation (welfare-enhancing effects) is greater than trade diversion (welfare-decreasing effects) depends on the structure of the economies to be integrated. The relative similarity of the European economies in 1957 led economists to vouch for economic integration through a common market given its plausible net positive effects in terms of trade creation (Pelkmans 2006).

The common market created by the EEC was the means to achieve an everlasting peace and prosperity on the continent. The common market has been a success in its realization. Indeed, from 1958 to 1972, while trade between the six founding Member States and the rest of the world had tripled, intra-community trade had been multiplied by 9 (Pelkmans 2006).

If the European integration had strong economic rationale due to the creation of a common market to the six founding Member States, the European common market could not last without a sufficient degree of institutionalization and a European law that encapsulate this economic rationale.

Law of European Integration: Institutions as Engines of Economic Integration

The effective working of the European integration requires a delegation of authority, even sovereignty, to supranational institutions. This requirement is derived both from the high reluctance of Member States to rely exclusively on other Member States for the implementation of the rules creating the common market and from the economic rationale of such institutions since decision-making costs are reduced, thanks to the specialization of supranational institutions.

The European Institutions

The institutions of the EEC are representatives of different interests. While intergovernmental institutions within which Member States play a key role represent national interests, supranational institutions within which European institutional actors are predominant represent the "community" interest.

The intergovernmental institutions of the EU are the Council of the European Union and the European Council. The supranational institutions of the EU are the European Parliament and the European Court of Justice.

The European Commission is the similar to a cabinet government, composed of 28 members of the Commission ("commissioners") – one per Member State – who execute the European Treaties. It represents the general interest of the EU. The President of the European Commission is proposed by the European Council and elected by the European Parliament.

The European Parliament is the directly elected legislative assembly of the EU since 1989 where European citizens are represented. Composed of 766 members, the European Parliament has incrementally gained legislative powers to achieve equal footing with the Council of the European Union under the ordinary legislative procedure generalized since the Maastricht Treaty of 1992. Under the ordinary legislative procedure provided at Article 294 of the Treaty on the Functioning of the European Union ("Lisbon Treaty" of 2009), the European Commission has the monopoly of legislative proposal. The European Parliament adopts a first position; if this position is approved by the Council, the piece of legislation is adopted. If the Council disapproves the Parliament's position, it adopts its own position and then communicates to the Parliament who either approves or disapproves the Council's position. In the latter case, a so-called conciliation committee is convened, composed of an equal number of Council's

members and of Parliament's members. If the conciliation committee adopts a joint text later approved by both the Council and the Parliament, the legislative act is finally adopted. If the conciliation committee fails to adopt a joint text, or adopts a joint text but not later approved, the legislative act is not adopted.

The European Court of Justice is designed to be the supreme court of the European Union by ensuring full compliance with European law by the Member States and European citizens. Seated in Luxembourg, the European Court of Justice is composed of independent judges from different Member States representing the different legal traditions present in Europe. Helped by the General Court and the Tribunal for Civil Servants, the European Court of Justice can hear cases either directly or from national tribunals (Turk 2010).

The Council of the European Union composed of national ministers represents the Member States with each Member States being allocated different voting weights according to its political power.

The European Council, originally in 1974 an informal of the head of states and governments of the Member States, has been recognized as a European institution in 1992. This high-level meeting instills the political impetus to the European Commission and, since the Lisbon Treaty of 2009, is chaired by the President of the European Union. The President is elected by the member of the European Council and its term lasts 2 years and half, renewable once.

This sophisticated institutional framework progressively, through treaty amendments, resembles one of a federation of states despite an added complexity due to the reluctance of Member States to renege upon their sovereignties.

The European institutions have been granted powers independent from Member States only to the extent that the Member States benefited from these delegations of powers in favor of these institutions acting as their agents. Indeed, the European institutions have been tasked with the completion of an internal market within the European Union. In order reach such result, the necessary functions have been delegated to the European institutions – this is the "neo-functionalism" theory as we shall discuss below.

The European Law

The completion of the internal market required European law to materialize the economic objectives of the European integration – namely, a customs union, the four economic freedoms, and a competition policy. European economic integration could only emerge through European legal integration. "Integration through law" has been the major force toward the completion of the internal market, with the European Court of Justice working as the primary engine of integration.

First, the European Court of Justice has elaborated a thought-provoking case law ensuring the greatest effectiveness of European law over national laws (Weiler 1991; Turk 2010). The European judges have granted to European law direct effect, making most of European law directly invokable by litigants before any national and European court. Also, the European judges have trumped national laws by continuously proclaiming the primacy of European law over any national laws (Weiler 1991). Moreover, the responsibility of Member States has been engaged any time European law is not respected in the given Member State by whatever national institution, citizen, or organization of that state. Also, the European Court of Justice reaffirmed its role as the supreme court of Europe by interpreting largely its monopoly of interpretation of European law and of annulling European law in light of the European Treaties (Weiler 1991). Therefore, not only have European judges confirmed the essential role of European law in the national legal orders, but also have they placed themselves as powerful judges in the most influential court in Europe (Bernard 2010).

Second, after having ensured its maximum effectiveness, European law has been developed and interpreted in order to maximize the extent to which the process of building the internal market is conveyed. Ernst Haas, one of the fathers of the functionalism theory, said that "the most inviting index of integration – because it can be verified statistically – is the economic one."

Market-building requires rules on the prohibition of discrimination since goods, services, capital, and labor must compete within a market on

a fair and open basis. The customs union is efficiency-enhancing on one part because it creates some trade but can also be efficiency-decreasing on the other part because trade can be diverted due to the discriminatory rule a customs union creates in favor of its members and in disfavor of the nonmembers. Hence, the customs duties (Article 30 TFEU) and "measures having an equivalent effect" (Article 34) are prohibited within the EU after the 12-year transitory period. The notion of "measures having an equivalent effect" has been given great scope by EU judges to encompass lots of regulatory measures. Indeed, "all trading rules enacted by Member States which are capable of hindering directly or indirectly, actually or potentially, intra-Community trade are to be considered as measures having equivalent effect to quantitative restrictions" said the EU judges in the famous *Dassonville* case of 1974. Despite the fact that some mandatory requirements can potentially be accepted as justification for some national regulations (see case Cassis de Dijon of 1979) above the justifications provided for in the treaties, the *Dassonville* formula has been continuously reaffirmed by the case law (Bernard 2010; Turk 2010).

Therefore, the EU legal order has been designed around the notion of competitive order – in other words, the internal market requires a law that ensures that competitive forces work openly, fairly, and on an equal basis for all market participants. Thus, in order to limit the reduction of economic efficiency and to increase the internal efficiency of the market, two legal dispositions had to be added at the top of the prohibition of customs duties and measures having an equivalent to quantitative restrictions. First, a nondiscrimination rule had to be an overarching principle of market-building. Second, a strong enforcement of a common competition law had to be ensured. The European successfully encapsulate both dispositions so that the internal market could be considered to have been completed by 1992 with the Maastricht Treaty.

First, the prohibition of discrimination on the basis of nationality enshrined in Article 6 of the Treaty of the European Union, in Article 21 of the Charter of Fundamental Rights of 2000 (a sort of EU "Bill of Rights"), ensures equality in the internal market, hence its smooth working. It is no surprise that the EU judges have interpreted the nondiscrimination principle in a very ambitious manner. The nondiscrimination principle applied very broadly in the field of the free movement of goods is progressively applied in a comprehensive manner to other economic freedoms. Hence, the "market access" has, in the eyes of the EU judges, to be secured in the broadest manner in all fields of the internal market through the resort of the nondiscrimination principle writ large (Bernard 2010). Market access must be ensured due to the benefits of transactional efficiency (in terms of minimization of transaction costs created by national regulations) expected to be reaped out by internal market participants (Portuese 2012, pp. 361–402).

Second, European competition law had preserved the internal market from obstacles to its achievement (the so-called objective of "market access"). The basic EU competition rules are found in the TFEU at Articles 101 and 102, which lay down a prohibition against restrictive agreements (cartels and collusions) and abuse of dominance, respectively. Articles 101 and 102 TFEU apply to practices or behavior that "may affect trade between Member States" (Toth 2008). The concept of "trade between Member States," defined in Commission Guidelines (No 2004/C 101/07 of 27 April 2004), implies that there must be cross-border economic activity involving at least two Member States – but, when trade between Member States might be affected, a restrictive agreement applied only in a single Member State will not be excluded from the scope of EU competition rules. Also, mergers are controlled according to Regulation 139/2004 of 2004. Finally, state aids are circumscribed in accordance to Article 107 of the TFEU. Entrusted with the task of investigating anticompetitive practices, the European Commission is invited to refer to the for potential cases to be tried. The high enforcement of European competition rules have participated in making the internal market a place where a fair and undistorted competition provides the economies efficiencies of a dynamic economy (Bernard 2010).

Overall, the removal of customs duties and charges having equivalent effect, the respect of the nondiscrimination and equality principle, and the strong enforcement of European competition rules have contributed to materialize European integration over decades. If European integration has been propelled by the European institutional actors such as the European Commission and the ECJ, it is also due to the particular appeal the pursuance of the notion of economic efficiency had on these supranational actors.

European Institutional Actors as Agents of Economic Efficiency: The Efficiency Hypothesis

The foundations of European law are substantiated in the general principles of EU Law. With the principle of equality and nondiscrimination discussed above, there are three main general principles that govern European law – the principle of subsidiarity, the principle of proportionality, and the principle of legal certainty.

The principle of subsidiarity is double-edged. Enshrined in the Treaty of the EU at Article 5 (3), it can be used to justify both increased centralization and increased decentralization. The principle of subsidiarity can justify further centralization because it requires that the most appropriate level of governance be chosen for exercising a particular power. A study of the principle in general, and in practice as interpreted by the EU judges, reveals that the principle of subsidiarity both is a principle of economic governance and is interpreted in an economically efficient way by the EU judges (Portuese 2011).

The principle of proportionality is a general principle of EU law explicitly stated in the EU Treaties nowadays at Article 5 (4). The EU proportionality principle can be divided in different sub-principles: (i) the review of the necessity of the measures to achieve the desired objective, (ii) the review of the suitability (or less-restrictive means test) of the measures for the achievement of the objective, and (iii) the review of the proportionality *stricto sensu*, whereby the burden imposed must be of proportion with the goal desired. If the first sub-principle is tantamount to an effectiveness criterion, the second sub-principle ensures the efficiency (ratio means-ends) of the measure, while the third sub-principle involves a balancing exercise of interests similar to a cost-benefit analysis which itself fosters Kaldor-Hicks efficiency. Therefore, the overall principle is and is interpreted as an efficiency principle (Portuese 2013).

Finally, the EU principle of legal certainty protects the legitimate expectation of market actors. Therefore, it ensures that changes in the law occur at a predictable path and in a reasonable way. If this guarantee is not respected, then damages must be awarded in order to compensate the loss of market actors' proprietary interests. This principle has been jurisprudentially created in the EU, so the European integration can happen without loss of proprietary interests, thus without disincentivized internal market actors (Portuese 2014a, b).

Along the general rules of European integration such as free movement rules and competitions rules, the general principles of EU law are efficiency-enhancing rules and have been interpreted as such consistently by the EU judges.

There is no coincidence however.

Indeed, one must recall the high distrust Member States had one another in the aftermath of the Second World War. Therefore, a relatively strong supranational institutionalization has been erected in order to eschew reciprocal and the tit-for-tat ill-fated strategies of classical treaties. Consequently, any political answers favoring one specific Member State or a given group of Member States at the expense of other Member States had to be seriously sidestepped. Indeed, welfare distribution would have been seen at that time as a means for a Member State to get a hold on the overall European integration. Thus, the European construction was only left, politically speaking, with a solution of welfare creation: an economic solution secured by the law (Portuese 2012).

In that prospect, economic efficiency is a rather neutral notion, despite its criticisms. It does not favor a specific Member State over another; it promotes the general interest of Europe inasmuch as supranational institutions were tasked to promote. Hence, supranational institutions of the EU have been keen to develop the economic

efficiency within the internal market because it is both an engine of integration and, very notably and interestingly for them, a politically accepted objective.

In light of the cursory economic analysis of European law discussed above, it can be said that there is tendency of European law in general and of European case law in particular to develop in sense that promotes economic efficiency. Therefore, a hypothesis of the economic efficiency of the European case law can be made, by reference to the classical and contested economic efficiency hypothesis of the common law (Portuese 2012).

If European integration has been in great majority about economic integration, the creation of a supranational polity having manifest powers over the economic environments of Member States but without possessing the democratic legitimacy of national democracies raises the everlasting concerns of European integration.

The Law and Politics of European Integration

If European integration has been for long driven by economic grounds unfolded by the ECJ case law (Turk 2010), the political integration was more diffident since Member States were both reluctant to share the prestige of their national legitimacies and to set up, from the onset or later at the end of twentieth century, a European polity capable of being compared to any form of federation or superstate (Weiler 2000). This reluctance can be understandable but is nevertheless problematic for at least two reasons: the very essence of European law becomes under fire given the low legitimacy of this "undemocratic" law (a), and the very essence of the European institutions also evolves into an easy target for criticisms due to the lack of the commonly recognizable features of national institutions shared by the European institutions (b).

In fact, the European Union is based on an inverted construction where market-building has preceded and primed over institutions building. The European Union, a sui generis organization, can be defined by the concept I introduce of *soft federation* (c).

Harmonizing Details and Enlarging High-Tail: Challenges to the Law

The European construction has developed European law to an extent that it becomes difficult to find sectors of the economy where regulations have not been harmonized at the European law. Indeed, the EU works predominantly as a regulator, privileging the normative power over military power.

The EU can only regulate, however, when it is directly related with the internal market. This is in accordance with the principle of conferral of powers.

When sectors are not harmonized, the so-called principle of mutual recognition applies: it means that the national rules have to be mutually recognized by the other Member States (Armstrong 2003). Hence, traders do not have to comply with extra rules that are purported to achieve objectives for which the rule of their home country has already been designed to achieve. Extra national regulatory costs are avoided unless mandatory requirements justify compliance with the host state rule. This process is often called "negative harmonization."

Therefore, as an incentive, the European institutions have been keen to harmonize at great pace. The removal of all barriers to trade can sometimes lead to the hastened removal of national rules. Therefore, harmonization with one single European rule is needed in order to foster the integration of the internal market. This process is often called "positive harmonization".

Negative harmonization is deregulatory; positive harmonization is re-regulatory.

The interesting point to raise in the following lines is not about how much, what, and when have we harmonized regulations in Europe. It is more about what consequences this process of continuous harmonization bears as challenge to the law as we understand it.

Coupled with the great pace of harmonization is the issue of enlargement (Weiler 2003). Not only have we integrated through harmonization with European rules throughout the European

construction, but also we have enlarged the EU from 6 founding Member States with 169 million inhabitants to 28 Member States with 500 million inhabitants, and it keeps enlarging!

These two major trends of European integration beg the question whether or not European law has enough democratic legitimacy in order to keep both harmonizing in its depth and enlarging in its ambit.

European regulations have recently been under greater scrutiny as for their costs and legitimacy, with the so-called Regulatory Impact Assessment ("RIA") that intends to balance the costs and benefits of each regulation. This tool has been developed under the framework of the "Better Regulation" program – the EU has to be a better regulator rather than a merely effective regulator. But the ambition of better regulating the EU leads to the very notion of legislative acceptability of EU law by citizens with respect to as EU law is perceived.

The law is traditionally understood as being a body of normative rules coherently forming a legal order and applied to a specific territory where subjects are legally binding themselves to this order with full consent. In other words, should the subjects change their mind in a democratically erected legal order, the institutions will be responsive and change the law. Here is the necessity of not binding one generation with another. That's why we have this expression "King in his Parliament": the rules in force are only those expressly or implicitly consented by the people to comply with.

But, as EU institutions can increasingly powers of economic and social regulations, the possibility for national democratically elected representatives to change the law applicable to their citizens. This argument does not work only for deregulating purposes: the mere changes in law are extremely difficult due to the decision-making process taking place at European level that leaves national representatives unpowered.

The notion of EU legal order faces multiple challenges not only for competing with national legal orders that bear high democratic legitimacy but also for ascertaining a minimum prestige owed by European citizens to their European legal order. This prestige is not superfluous, it is fundamental: a law without prestige is no law inasmuch as a law without sanction is no law. European law with little prestige hardly is law.

National Democracies and Supranational Governance: Challenges to the Politeia

If European law faces great challenges in terms of its acceptance and position with respect to the perception of national legal orders, it is also because European integration, again, has been predominantly about economic integration.

The failure of planning, the inability to foresee, the absence of courage, and the reluctance to go supranational have impeded the necessary process of polity-building required by any market-building process. Indeed, the *Politeia* (the Greek word for "polity") defines a specific form of government into which citizens can believe be part of. It does not necessarily mean "democratic" as contemporary understanding evoke; neither does it evolves the notion of Republic nor of nation-state. It only refers to the requirement of having clear and intelligible institutions that citizens recognize as legitimate. Unfortunately, despite efforts made by the Member States and the European institutions to "democratize" the EU by building a supranational polity, the EU institutions are still perceived as far from European citizens. Indeed, the organization since 1979 of direct elections for the European Parliament has not helped this core institution to get credit or to avoid lots of abstentions among voters who are European citizens. Also, the recent creation, with the Lisbon Treaty, of the President of Europe elected by the European Council has not helped European citizens to identify Herman Van Rompuy, the first European President, nor did it helped European citizens to feel he is their European representative (Weiler 2000).

Lots of academic debates are of little help when it comes to increasing the democratic legitimacy of European institutions and to focusing more on polity-building rather than on market-building. One trend of the literature, new functionalism, has argued that the step-by-step integration would "spill over" continuously so that, eventually, an integrated and federated Europe would come to the fore. Within that process, the rise of

a political Europe would emerge as economic integration cannot last without institutional integration (Haas 2004; Sandholtz and Stone-Sweet 1998; Stone Sweet and Fligstein 2001). This process is triggered principally by supranational institutions who act with great independence from Member States. Another trend of literature, labeled "liberal intergovernmentalism," argues contrariwise that Member States have retained, throughout the European construction, considerable latitude over the European institutions' powers (Moravcsik 1993). The dynamics of European integration has been, according to that perspective, aligned with the powerful Member States' interests who succeeded in deepening European integration while preserving their powers.

Beyond this academic debate lays the discussion on a polity-building at the European level which could, if achieved successfully, deliver to the market-building process an enhanced democratic legitimacy. The failure to envision such polity-building or the reluctance to achieve it can be explained by the very nature of the EU. The EU actually is, as I coin, a "soft federation."

The Political Economy of the Current European Integration: The Soft Federation Hypothesis

In law, soft law is a well-known expression used to designate some quasi-normative rules that are neither hard law because they lack the enforceability of legal rules nor no law because they carry some normative guidelines of certain legal valence.

This concept of soft law can be extrapolated to political sciences and defines a federation that is both insufficiently powerful to be designated as a "hard federation" and insufficiently powerless to be designated as a confederation.

The establishment of a federation generally comes with both the exclusive recognizance of the federal state as depository of sovereignty (both internally and externally), with the creation of police and military powers, the exclusive power over monetary powers, and a preeminence over budgetary powers, and the exclusive power to carry out diplomatic activities. Contrariwise, the federation can intervene less often and deeply on economic and social matters, the setting up of normative standards, and the policy interventions in fields such as education and research, environmental protection, heath policy, etc. But on the other hand, the EU can no longer be considered as a confederation since majority voting has been generalized, whereas confederation are defined by unanimity voting in order to respect each Member State's sovereignty.

Therefore, an inverse institutional architecture appears compared to classic federation framework: the most integrated parts within the EU are the least one in a federation, while the least integrated parts of the EU are at the essence of the federal level. Therefore, the EU is a federation without the common powers of a federation – it is a "soft federation."

This hybrid situation is not a middle-ground situation: it may very well last for long, and hence, the "soft federation" format may become a new form of government rather than a temporary one if the "soft federation" hypothesis is confirmed over time.

Conclusion

European integration has been a great success in terms of enhancing the economic efficiency through the process of market-building, while it has remained a great disillusionment in terms of enhancing democratic legitimacy through the process of polity-building.

The two major trends identified in European integration lead to formulate a dual hypothesis: an efficiency hypothesis whereby supranational actors push for internal market rules to promote economic efficiency and a "soft federation" hypothesis whereby Member States push for the emergence of a peculiar federation which protect the core of national sovereignties.

Overall, European integration is a unique process successful enough to witness trials of replications throughout the world (e.g., African Union, Union of South American Nations, Arab League, etc.). This is certainly the strongest proxy for evaluating the results of 60 years of European integration.

References

Bernard C (2010) The substantive law of the EU: the four freedoms, 3rd edn. Oxford University Press, Oxford

Duchene F (1994) Jean Monnet. The first statesman of interdependence. Norton, New York

Haas EB (2004) Introduction: institutionalisation or constructivism? In: The uniting of Europe: political, social and economic forces 1950–1957, 3rd edn. University of Notre Dame Press, Notre Dame, pp xiii–lvi

Kant I (1992) Perpetual peace. A philosophical essay. Thoemmes Press, Bristol

Montesquieu B (1758) Oeuvres Complètes: De l'Esprit des Lois (Complete works: spirit of the laws). Gallimard/Pléiade, Paris (1951)

Moravcsik A (1993) Preferences and power in the European community: a liberal intergovernmentalist approach. J Common Mark Stud 31:473–524

Pelkmans J (2006) European integration, methods and economic analysis. Pearson Education/Financial Times, Harlow/Prentice Hall

Portuese A (2011) The principle of subsidiarity as a principle of economic efficiency. Columbia J Eur Law 17:231

Portuese A (2012) The principle of economic efficiency in the ECJ case-law (Le Principe d'Efficience Economique dans la Jurisprudence Européenne). Thèse de l'Université Paris II Panthéon-Assas, Paris. Accessible à: https://docassas.u-paris2.fr/nuxeo/site/esupversions/c51d2b4b-4981-4c62-8531-285d0804dc05

Portuese A (2013) Principle of proportionality as principle of economic efficiency. Eur Law J 19:612–635

Portuese A (2014a) The principle of legal certainty as a principle of economic efficiency. Eur J Law Econ. https://doi.org/10.1007/s10657-014-9435-2

Portuese A (2014b) The case for a principled approach to law and economics: efficiency analysis and general principles of EU law. In: Mathis K (ed) Law and economics in Europe, Foundations and applications. Springer Books, Dordrecht, pp 275–328

Stone Sweet A, Fligstein N (2001) The institutionalisation of Europe. Oxford University Press, Oxford

Toth AG (2008) The Oxford encyclopaedia of European community law, vol II: The law of the internal market (2005); vol III: Competition law and policy. Oxford University Press, Oxford

Turk A (2010) Judicial review in the EU. Elgar Publishing, London

Weiler JHH (1991) The transformation of Europe. Yale Law J 100:2403–2483

Weiler JHH (2000) Epilogue The Fischer Debate In: Joerges C, Mény Y, Weiler JHH (eds) What kind of constitution for what kind of polity?: responses to Joschka Fischer. Robert Schuman Centre, EUI

Weiler JHH (2003) In: Begg I, Peterson J, Weiler JHH (eds) Integration in an expanding European Union: reassessing the fundamentals. Blackwell Publishing, London

Further Reading

Armstrong KA (2003) Governance and the single European market. In: Craig P (ed) The evolution of EU law. Oxford University Press, Oxford, pp 745–789

Baldwin R, Wyplosz C (2009) Economics of European integration, 3 rev edn. McGraw Hill Higher Education, London

Cecchini P (1988) The European challenge 1992. Wildwood House, Aldershot

CEPR (1993) Making sense of subsidiarity: how much centralisation for Europe? London (study group Begg et al)

CEPR (2003) Built to last: a political architecture for Europe. London

Cohn-Bendit D, Verhofstadt G (2012) For Europe! Manifesto for a postnational revolution in Europe. Carl Hanser Verlag, München

Craig P, De Burca G (2011) EU Law. Text, cases and materials, 5th edn. Oxford University Press, Oxford

Diez T, Wiener A (2004) European integration theory. Oxford University Press, Oxford

Eeckhout P (2005) External relations of the European Union: legal and constitutional foundations. Oxford University Press, Oxford

Gilbert M (2012) European integration: a concise history. Rowman & Littlefield, London

Hix S (2005) The political system of the European Union, 2nd edn. Palgrave, London

Holle W (2006) The economics of European integration: theory, practice, policy. Ashgate, London

Howells G (2002) Federalism in USA and EC – the scope for harmonised legislative activity compared. ERPL 10:601–622

Olsen JP (2001) Organizing European institutions of governance: a prelude to an institutional account of political integration. In: Wallace H (ed) Interlocking dimensions of European integration. Palgrave, London

Pelkmans J (2002) Mutual recognition in goods and services: an economic perspective. College of Europe, Bruges. BEEP briefings no 2, December. www.coleurop.be

Pelkmans J, Labory S, Majone G (2000) Better EU regulatory quality: assessing current initiatives and new proposals. In: Galli G, Pelkmans J (eds) Regulatory reform and competitiveness in Europe, vol 1. E. Elgar, Cheltenham

Sandholtz W, Stone-Sweet A (1998) European integration and supranational governance. Oxford University Press, Oxford/New York

Smith A, Venables A (1991) Economic integration and market access. Eur Econ Rev 35:388–395

Weatherill S (1995) Law and integration in the European Union. Clarendon, London

Weatherhill S (2005) Harmonisation: how much, how little. EBLR 533

Woods L (2004) Free movement of goods and services within the European community. Ashgate, London

European Law

► European Community Law

European Nationality

Paul-Ludwig Weinacht
Güntersleben, Germany
Würzburg University, Würzburg, Germany

Synonyms

EU citizenship; European identity; Europeanization

Definition

European Nationality (EN) is a concept with two meanings: the present legal, political, and cultural reality of the EU ("actual real concept") and the possible futures of the EU ("target oriented concept"). Both sides of the concept, the "actual real" and "target oriented" ones, are employed in the ongoing process of enlargement and consolidation processes of the 28 (27) EU member states.

An Actual Concept

As an *actual concept* of our legal reality, EN means the EU-citizenship which since 1991 is tied to the primary national citizenship of an EU member state as a secondary one. In the treaty of Maastricht, all citizens of EU member states received the status of general and equal EU-citizenship (Art 17 sq EC). It does not replace the national citizenship, who's lending or withdrawal remains in the sovereignty of member states – however it complements it in several aspects. The EU-citizenship offers firstly the freedom to settle down and to establish oneself in all member states (Art. 18 EC); secondly it gives the right to vote in local elections of the state in which someone is legally settled and the right to vote for the European Parliament (Art. 19 EC). Thirdly, the right of diplomatic and consular protection (Art. 21 EC), the right of petition, the right to turn in writing to the organs of the community, p. e. to the representative of the citizens, and to get an answer in the chosen language (Art. 21 EC).

The program by the EP "Europe for the citizens" (2007–2013) intends "to integrate the citizens better in the process of the European unification"; it promotes "the co-operation between the citizens and their organisations from different nations." The Single European Market is the most integrated institution in Europe. Here the famous "four freedoms" are granted since 1987: free traffic of persons, of services, of goods, of capital. These freedoms are limiting more and more the sovereign action of member states. They act – if necessary by the European Court – as engines for the strengthening of the political system of the EU.

Concerning the political and sociocultural meaning of EN, the EU- citizenship with its rights in annex is largely pre-determined by national preconditions. Conflicts in the context with the banks crisis and in the Brexit-movement in Great Britain showed it. Until now, European citizens as a whole do not form a nation. They do not fulfil the criteria of a nation: neither in the sense of the German or slave concept of nation (origin, history, language, etc.) nor in the French understanding of the word (support for the constitution of one and un-dividable republic).

The EU-citizens in their respective states have a pragmatic relation to the question of more intensive adherence to their nation or to Europe. The answers measured by the Eurobarometer-Survey are often: adherent to collective identities. This refers to "multiples" collective identities in the EU (Trenz et al. 2003, p 17). The mind of the citizen is influenced by the preparation and implementation of the election to the EP and – independently from the election campaign – intensified by culture programs of the EC (2007–2013). "Integration" and "participation" are headwords for the construction of a

well-united European Community, whose identity is based on commonly accepted values (http://europa.eu/legislation).

A Target-Oriented Concept

As *a target-oriented concept,* EN is precarious in so far as it applies to the controversial project of a "sovereign European state." Pro-European politicians like to symbolize the way to this kind of state by a dynamic concept which embraces different fields: "Europeanization." If the European citizenship would accept criteria of a sovereign state, it would be superior to the national citizenship. In the face of a sovereign position of the EU, some authors mean that common values could be effective in favor of the enlargement of European politics. A specialist means that strong structures and a dense net of competences in European institutions favor the "European identity and public spheres" (Risse 2010).

Prototypical forms of nation-building could serve as an historical example: the French nation as a "product of the state," because its origin is a sovereign monarchic rule; the German nation as "the base of the state," because it was – as a type of "community of cultural origin" (Weinacht 1994) – historically prior to the state (Brubaker 1994). The idea of Europeanization could be discussed too at the historic example of England: The political institutions created the English nation, and the political institutions were legitimated by the national English culture (Schulze 2004). It is questionable, if the insinuated lines of nation and institution building will result in a clear definite idea of state and if such an objective will be generally accepted. It is not without reason that the EU-Commission avoids fixing its institutional targets on a definite idea of state ("finality") and speaks about "an ever closer union among the peoples of Europe" (Treatise of Maastricht 1992/93).

The "Europe of citizens," which is intended, grows for the time being beside the nations in the form of cross-border sectors of a civil society, which develop both pro-European and anti-European tendencies. Supporters of this phenomenon try to see in a European "active society of citizen" a new model (Münch 1999).

Is EN Desirable?

Concerning the desirability of European nationality, the Treaty of Maastricht (1992/93) defines the target of the EU as an "ever closer union among the peoples of Europe." However some Eastern European ("Visegrad") States have lost this consensus and tend – like France under de Gaulle ("Europe des patries") – to a "Europe of nations." With the referendum from June 2016, Great Britain cancelled the treaty of Maastricht in favor of an independent future for its nations. However, the aftereffects of Brexit are controversial. Britons who feel and wish to be part of the European project will loose with European citizenship the privileges inherent in it (especially "freedom of movement"). Many of them have begun to apply for the citizenship of their host country. The Irish embassy in London received 8017 applications for Irish nationality, compared to only 689 applications in 2015. The reason was not the search for a new identity, but the desire to maintain security of residence and the right to travel and live in the 27 countries that will remain members of the EU after the UK leaves (Stone 2017). Politicians and journalists made proposals in favor of a European associate citizenship "for nationals of a former member state"(Henley 2016; Stone 2016).

A European Citizens' Initiative (ECI) came out informally known as "Flock Brexit" with a Facebook group titled "EU Citizenship for Europeans: United in Diversity in Spite of jus soli and jus sanguineous." It demonstrates not only the pure pragmatic character of EU Citizenship but Citizenship without political Membership. Although political rights are demanded sometimes, European citizenship is no more than a title associated with the citizenship of an EU member state.

In legal terms, which are underlying the status of European citizen, EN is an actual concept. Its

actual value remains in transnational civil rights, especially the right of movement and entitlement to the same treatment including welfare rights, as the citizens of the country of residence. Prudent politicians declare that the European Union does not substitute the sovereign nations but gives them forward compatibility. That's why EN would remain for a long period a secondary identity.

The transition from the formal status of EU-citizen to a substantial European nationality, realizing what the concept oriented to targets means, is in a quite distant future.

Cross-References

- ▶ European Community Law
- ▶ European Integration

References

Brubaker R (1994) Citizenship and nationhood in France and Germany. Harvard University Press, Cambridge, MA

Henley J (2016) EU citizenship proposal could guarantee rights in Europe after Brexit. The Guardian, 8 November

Münch R (1999) The transformation of citizenship in the global age. In: Kriesi H et al (eds) Nation and national identity. Verlag Rüegg, Chur/Zürich, pp 109–125

Risse T (2010) A community of Europeans? Cornell University Press, Ithaca/London

Schulze H (2004) States, nations and nationalism: from the Middle Ages to the present. Blackwell, Oxford

Stone J (2016) How to keep your EU citizenship after Brexit, British nationals are set to lose out, but there are some workarounds, 14 June, BST

Stone J (2017) Brexit vote creates surge in EU citizenship applications. The Guardian, 3 October

Trenz H-J et al (2003) Demokratie-, Öffentlichkeits- und Identitätsdefizite in der EU. In: Klein A, Koopmans R, Trenz H-J et al (eds) Bürgerschaft, Öffentlichkeit und Demokratie in Europa. Leske und Budrich, Opladen pp 7–19

Weinacht P-L (1994) Nation als Integral moderner Gesellschaft. In: Gebhardt J, Schmalz-Bruns R (eds) Demokratie, Verfassung und Nation. Die politische Integration moderner Gesellschaft. Nomos Verl.-Ges, Baden-Baden, pp 102–122

European Patent System

Bruno van Pottelsberghe de la Potterie
Solvay Brussels School of Economics and Management, Université libre de Bruxelles, Brussels, Belgium

Definition

The European patent system is defined as the policy mechanisms, jurisdictions, and institutions in Europe which allow inventors to acquire and enforce industrial property rights over their inventions.

Introduction

The European patent system is a terminology which is often oversimplified and described through the European Patent Convention (EPC). This is a multilateral treaty signed in October 1973 which essentially consists in creating the European Patent Organisation and in providing a legal system for the granting of European Patents. EPC patents, or so-called "European" Patents, are actually not really European, in the sense that they have to be translated and validated in national patent offices in order to be enforceable in national jurisdictions. The European patent system is more complex than the EPC, as it includes three layers of legal rights (national patents, the current European Patent, and the forthcoming Unitary Patent). Two types of patent offices grant these rights: national patent offices (which grant national patents) and the European Patent Office for the granting process of European and Unitary Patents. Litigations currently take place at the national level, and for unitary patents patent litigations should be treated by a Unified Patent Court, whose design is currently being framed and negotiated, as of January 2015. This entry aims at providing a wide understanding of the European patent system, its strengths, its weaknesses, and its challenges.

Chronology of Three Patent "Layers"

As of the late nineteenth century, all European countries had a national patent office that would grant national *patents*. The broad patentability conditions, found in most patent systems around the world, include novelty and inventiveness. In order to be granted, a patent must be novel (i.e., it should not have been published or made public prior to the application), and it must be inventive (there must be an inventive step, or non-obviousness, with respect to the prior art). If the patent owner aimed at extending its geographical coverage, it had to file subsequent applications in other European countries. This was particularly complex and costly, as it led to additional translation costs, application fees, and examination fees, not to mention renewal fees to maintain the patent in force once it was granted, for a maximum duration of 20 years from the first date of application. According to the Paris Convention (March 1883), which is still in force today, patent owners have maximum 1 year after their first application to extend their patent abroad while not being refused on the ground of the novelty condition (a patent published elsewhere would constitute a prior art) (Table 1).

A second layer of patent protection was activated in 1978. It is known as the "European Patent" and was set up through the creation of the EPC, signed by 38 countries, as of June 2014 (including all countries of the European Union).

European Patent System, Table 1 Patent applications by patent office and origin, 2013

	Applications		Patents granted	Patents in force	Cumulated fees (4Y)
Name	Total	% nonresident (%)	Total	Total	EUR
Austria	2,406	10	1,256	110,202	830
Belgium	876	18	745	..	485
Bulgaria	297	5	125	1,431	n.a.
Croatia	253	9	159	4,243	n.a.
Czech Republic	1,081	9	611	7,780	n.a.
Denmark	1,534	13	309	51,277	1,398
Finland	1,737	8	711	47,058	1,890
France	16,886	13	11,405	500,114	728
Germany	63,167	25	13,858	569,340	988
Greece	717	3	282	2,966	513
Hungary	708	9	1,351	5,237	807
Ireland	390	15	214	108,218	700
Italy	9,212	10	8,114	68,000	1,160
Luxembourg	169	33	112	20,421	334
Netherlands	2,764	16	2,029	12,704	252
Poland	4,411	4	2,804	47,610	235
Portugal	669	3	130	36,782	200
Romania	1,046	5	451	17,100	550
Slovakia	210	12	115	2,755	n.a.
Spain	3,244	7	3,004	36,893	510
Sweden	2,495	7	685	14,539	658
UK	22,938	35	5,235	469,941	229
Total national offices (EU)	**132,921**	**22**	**49,056**	**2,134,611**	**n.r.**
European Patent office	**147,987**	**50**	**66,696**	**n.r.**	**4,309**

Source: figures on patent applications, granted and in force, are from WIPO statistical series released in 2014. Fees are from de Rassenfosse and van Pottelsberghe (2013) and represent cumulated fees from application to grant, in EUR

In short, once granted by the European Patent Office, a patent must be managed and enforced at the national level. This "European" terminology is therefore often confusing, and certainly inadequate, as the "European" dimension only takes place for the processing of the search report and the substantive examination of patent applications. They are centrally granted by the European Patent Office. Once granted these patents must still be validated and renewed in the national patent offices that fit with the desired market coverage of the patent owner. The cost of a European Patent is therefore relatively high and depends on its geographical coverage, as inventors must pay their validation fees and several other costs (intermediates, patent attorneys, renewal fees). An opposition procedure before the EPO can be initiated by any person or institution during a period of 9 months after the date of decision to grant a patent. Any subsequent litigation is operated at the national level (see Guellec and van Pottelsberghe 2007 for a broad introduction to the economics of the European patent system).

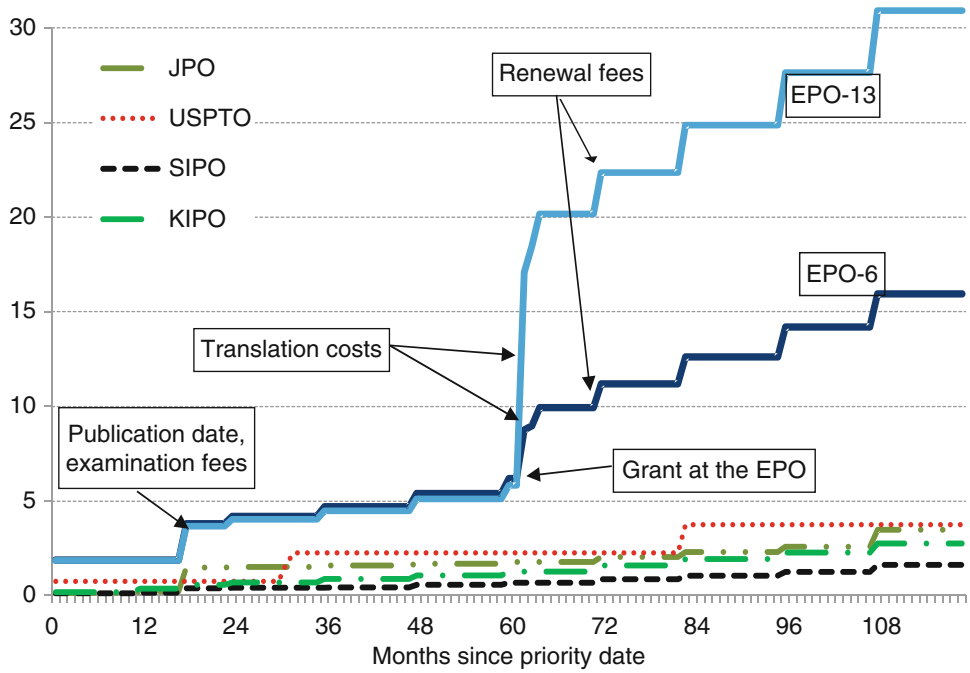
Table 1 presents these first two "layers" of patent protection. National patent offices receive national patent applications (ranging from a few hundreds in small countries to more than 60,000 in Germany and 23,000 in the UK). The cumulated number of patent applications at national patent offices (about 132,000) is smaller than the 150,000 patent applications filed at the EPO. Interestingly, the EPO is particularly targeted by foreign companies, as 51% of its patent applications are filed by non-EU residents. National patent offices are less targeted by foreign companies (only 21% of patent applications are filed by nonresidents).

Countries inside Europe have adopted different fee structures and policies. The last column of Table 1 shows that the cumulated fees from application to grant fluctuate from less than 300 EUR in the UK or the Netherlands to more than 1,000 EUR in Italy, Finland, and Denmark. When it comes to the European Patent, the fees for patent prosecution before the EPO are much higher than in national patent offices, with more than 4,000 EUR of cumulated fees, from application to grant. And these cumulated costs do not reflect the maintenance rates of European Patents in each national jurisdiction.

The fragmentation of the European market is ultimately illustrated by the very strong heterogeneity in the number of patent in force across countries. Germany, France, and the UK have by far the largest number of patents in force, more than 450,000 (be they granted by the EPO or by their national patent offices). All other EU countries have much less than 100,000 patents in force, due to their smaller size and hence lower market attractiveness. It is worth noticing that a European Patent becomes particularly expensive once it is granted, due to the translation costs and cumulated renewal fees. This is well illustrated in Fig. 1, which shows the evolution of cumulated costs for patents that aim at being put in force in 6 or in 13 European countries, as compared with Chinese, Japanese, South Korean, or US patents.

The main consequences of the system that combines national and "European" Patents are described in Mejer and van Pottelsberghe (2012) and in van Pottelsberghe (2010). The most visible one is related to the prohibitive costs of patent protection in Europe, as compared to several other regions in the world. A European Patent with protection secured for 10 years in 13 countries will cost more than 30,000 EUR, against less than 5,000 in the USA, Canada, China, Brazil, Japan, or South Korea.

Another important consequence of such fragmentation is related to the broad legal uncertainty associated with the European Patent, especially for technology-based spin-offs. Legal uncertainty is the outcome of a dual system in which the EPO grants patents centrally but where national patent systems have the ultimate power to validate, invalidate, and assess infringement proceedings relevant to their own jurisdiction. European Patents which are particularly valuable have a high propensity to be litigated, in several countries. Parallel litigations are extremely expensive and time-consuming, especially for small entities. Then come the incongruities, whereby a patent can be maintained valid in one jurisdiction and invalidated in another. And even if a patent is maintained valid in several jurisdictions, they



European Patent System, Fig. 1 Ten-year cumulated patent fees for a European Patent (validated in 6 and 13 countries), compared with the US and three Asian countries (Data source: Mejer and van Pottelsberghe (2011); cumulated costs (in Euro) from application to grant and renewal in 6 or 13 EU countries (for the EPO) and in the USPTO (USA), KIPO (South Korea), JPO (Japan), and SIPO (China))

may reach opposite conclusions regarding infringement proceedings. Parallel imports are also relatively easy, as infringing products might enter Europe through a country with no patent being enforced and then easily being distributed over Europe. An additional source of inconsistency deserves attention: the fact that national patent offices may grant a national patent even if it has not been applied for at the EPO or if it has been refused by the EPO. Procedurally, it is perfectly permissible to make simultaneous filings at one or several national patent offices and at the EPO. In other words, the granting process orchestrated by the EPO can be "bypassed" if one or more applications are made directly to national patent offices. This practice may have a number of explanations, some innocent (the applicant being interested by only one or two national markets), some less so (a perception that some national offices are a "soft touch" for applications compared with the EPO, and hence applications in the gray zone regarding their patentability can legitimately be filed at the central and national levels).

These prohibitive costs and legal uncertainty are the main reasons which have prompted policy-makers, for more than 50 years, to try to create a truly European Patent, first called "Community" patent and later entitled "Unitary" Patent, a patent that would bring protection over the whole European Union. This *third layer* has been voted in 2013 at the EU Competitiveness Council at the Ministerial level (Regulation EU No. 1257/2012 for the introduction of the European Patent with unitary effect or the Unitary Patent and Regulation EU No. 1260/2012 for the translation arrangements for the Unitary Patent) and ratified by 25 countries (Croatia, Italy, and Spain did not sign the treaty). The regulations have been adopted by way of enhanced cooperation – a tool that allows a group of EU member states to go ahead with integration and legislative projects when the required majority in the Council cannot be achieved; it is used only as

a last resort. It consists in creating a Unitary Patent, which would be automatically valid in the 25 member states of the EU, at the request of the patent owner and once the patent is granted by the EPO. The Unitary Patent is therefore a new option for the holder of a European Patent, who will choose between the classical "European" route of selecting the desired states for protection and a unitary protection covering 25 member states. During a transitional period of 12 years, the patents granted in German or French will have to be translated into English, and the patents granted in English will have to be translated into German and French. The translations will not have any legal effect; they are for information purposes only.

In a nutshell, the legal architecture of the Unitary Patent includes the "Unitary Patent Regulation" which consists in the introduction of the European Patent with unitary effect. Then comes the "Translation Regulation," which describes the translation arrangements for the European Patent with unitary effect. These two regulations will be applicable for a maximum of 25 EU member states only on the date of entry into force of the Agreement on a Unified Patent Court (UPC) and only in the countries which will ratify it. As of January 2015, the Unitary Patent is therefore far from being operational, several details of its implementation having still to be framed and negotiated, and the number of countries it will actually cover is still unknown. These details are the medium-term challenge of the European patent system.

Medium-Term Challenge: Designing the Right System

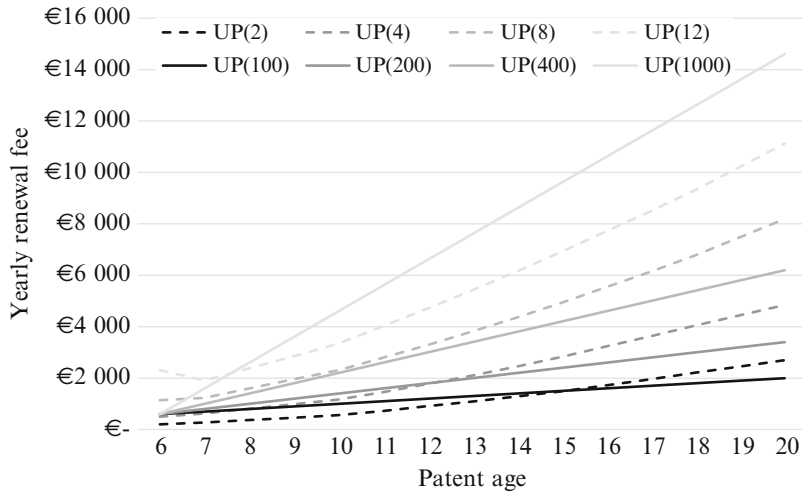
"Challenge" is an appropriate wording, because the decisions that will be taken by policy-makers, hopefully before June 2015, will substantially affect the use of the Unitary Patent and the broad effectiveness of the European patent system. Two important facets have to be framed: the level and structure of renewal fees and the setting up of a centralized litigation system, orchestrated by the Unified Patent Court.

Renewal Fees. . . For Who?

The Unitary Patent renewal fee structure has been the subject of intense negotiations for several years. This is a highly sensitive issue, as it has financial consequences for both the EPO and the national patent offices (NPOs), as well as for all the stakeholders of the European patent system: inventors, patent attorneys, lawyers, or translators. The renewal fees paid to NPOs by the owners of European Patents are split into two equal parts. Half the amount stays in the NPOs budget, and the other half is transferred to the EPO (cf. Danguy and van Pottelsberghe 2011a, b, 2014). For the NPOs of large frequently targeted countries, the patents granted by the European Patent Office generate massive resources (more than EUR 100 million in Germany and about EUR 50 million for France and the UK), equivalent to several times the annual working budget of their NPO. Even for smaller countries, the income generated by the renewal fees of European Patents exceeds by far the budget of their NPO induced by the proceeding of national patents. With the Unitary Patent, renewal fees would be collected by the EPO, and half the amount would be distributed across NPOs. The "renewal fee" debate is therefore not only related to the level these fees, but as well to the distribution key (what percentage will be allocated to each country, according to which criterion?). Some NPOs defend the position of setting up very high renewal fees, as they expect it will have a minor effect on their income (most patent owners would opt for the current "European" route, and those who will opt for the "Unitary" route will generate more resource).

In summary, as of January 2015 a self-fulfilling prophecy is in the air, whereby the decision taken by policy-makers on the renewal fees of the forthcoming Unitary Patent will affect its success and its perceived effectiveness (cf. Danguy and van Pottelsberghe 2014).

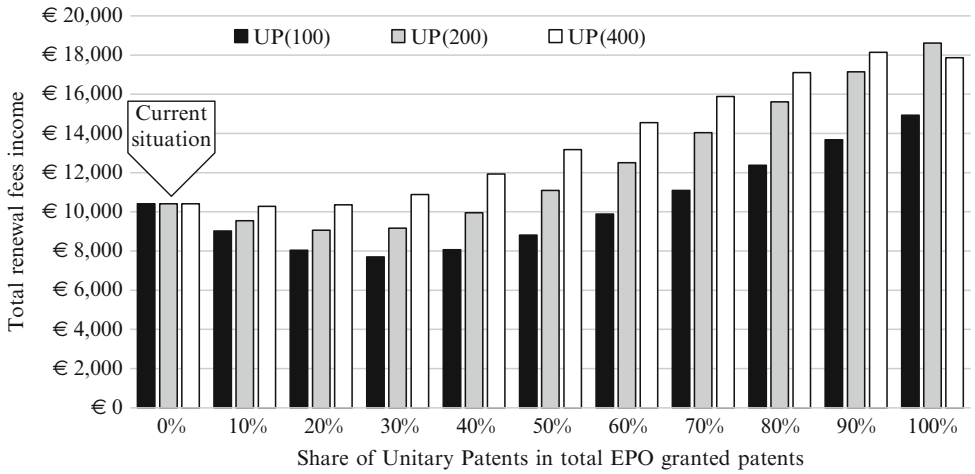
Two renewal fee schedules can be considered: (1) summing up the current European Patent fees of several countries, for instance, those in which granted patents are most frequently validated, like Germany, France, the UK, etc., and (2) adding a fixed increment each additional year of the patent



European Patent System, Fig. 2 Potential renewal fee schedules for the Unitary Patent (Note: *Dashed lines* show fee structures which sum the fees of X countries that are the most frequently targeted for protection. *Plain lines* show the fee structures with fixed annual increment (described between parentheses, starting fee at EUR 600 on year 6, and then an increment of 100 EUR (200 EUR, 400 EUR, or 1,000 EUR) each additional year. Source: Danguy and van Pottelsberghe (2014)))

life. Figure 2 illustrates four alternatives for the former structure – the sum of fees of the first 2 (4, 8, or 12) countries – and four alternatives for the latter one. The UP(200) fee schedule could be considered as the most appropriate because it is simpler than the additive fee structures, and it corresponds to what the business sector is currently paying. Indeed, van Pottelsberghe and van Zeebroeck (2008) showed that the average geographical scope of protection is about four countries – UP(4) is similar to UP(200) – for the patents granted 15 years ago by the EPO. With UP(200), an applicant would pay cumulated renewal fees of about 15,000 EUR to keep its patent enforced for 20 years in the 25 EU member states. This absolute cost is affordable in relative terms, given the large geographical scope of protection provided by the Unitary Patent. The EUR 30,000 figure is similar to what a patent holder must be ready to pay in the current European patent system for less than 10 years of protection in 13 countries.

But the important question from the patent offices' viewpoint is whether the Unitary Patent would generate the same resources with its central renewal fees for the EPO and for all NPOs. Danguy and van Pottelsberghe (2014) provide some answer, by simulating the net present value of renewal fee income for the average patents granted by the EPO, within the current system and within a two-layers system. Figure 3 shows the sensitivity of these simulations to different Unitary Patent fee schedules and to the expected share of Unitary Patents in total EPO-granted patents. Two observations might be drawn from these simulations. First, the higher the level of Unitary Patent renewal fees, the higher the total renewal fee income per average patent granted by the EPO, independently from the share of Unitary Patent. In other words there is a natural temptation to set high renewal fees. Second, with low Unitary Patent renewal fees, there is a U-shaped relationship between the share of UP and the total renewal fee income per average patent. This is due to the substitution effect between the two types of patent. The total renewal fee income collected by patent offices could actually be lower than the income in the current situation if the unitary patent system is not attractive enough (i.e., low share of UP) and has low renewal fees. At first glance these simulation results strengthen the argument for high UP renewal fees, so that they generate more



European Patent System, Fig. 3 Simulated total renewal fee income per average patent (Note: the *bars* represent the net present value of a patent granted by the EPO and which would opt for the Unitary Patent route, taking into account its expected maintenance rate. Source: Danguy and van Pottelsberghe (2014))

income than under the current situation. However, very high fees would lead to a low use of the new unitary patent.

But even with aggregate renewal fee income stable or larger than the current system, patent offices logically worry about the share of this income that will be allocated to them. Danguy and van Pottelsberghe (2014) have simulated the revenue stream of national patent offices under the dual system. Their results show that only a handful of NPOs could be negatively impacted by the creation of the unitary patent system. Except for Germany, the budgetary losses for these patent offices are very low and only occur with relatively low shares of Unitary Patents in total EPO-granted patents. Improving the revenue prospects could also be achieved if the number of patents granted by the EPO increase and if national renewal fees would increase.

Which Design for the Unified Patent Court?

The second medium-term challenge relates to the design of the forthcoming Unified Patent Court (UPC), aiming at proceeding unitary patent litigations in Europe. In order to come into effect, the UPC Agreement must be ratified by at least 13 member states (including Germany, France, and the UK). A logical consequence is that the Unitary Patent will cover only those member states where the UPC Agreement is in force. There is therefore a high probability that the Unitary Patent starts with much less than 25 member states, because converging to a common patent court denominator might prove to be complex and lengthy.

Indeed, the current national litigation systems are highly heterogeneous, reflecting important differences across national jurisdictions. For instance, some countries have technically qualified judges and others not, and the wage of judges is particularly high in the UK. Total litigation costs vary significantly across countries. The UK is by far the most expensive jurisdiction, with costs that are nearly as high as the cumulated costs in France, Germany, and the Netherlands. In case of multiple parallel litigations across jurisdictions, cumulated costs vary from 310 thousands euros before the four courts of first instance up to 3.6 million euros when taking account of the cost of appeal at second instance. Multiple litigations are particularly prohibitive in Europe, especially for individuals and SMEs, and can be more than twice as high as in the USA (see Mejer and van Pottelsberghe 2012). The challenges here will be to find a proper balance and a design that rally a majority of national litigation systems.

It is the responsibility of the Contracting States to set up the Unified Patent Court. A Preparatory Committee has been created in September 2013 to set out a road map for the establishment and coming into operation of the court. It has five main working areas, including the legal dimension, the financial sustainability, the human resource and training components, the design of the information system, and the facilities. It seems highly probable that there will be a central division of the Unified Patent Courts, with three Divisions, specialized in specific technological areas, and based in Paris, London, and Munich. Then, there will be regional divisions and local divisions, covering one to several countries, depending on the geographical areas and country size. The envisaged local divisions and their working language between parentheses would be based in "England and Wales" (English), the Netherlands (Dutch and English), France (French), Germany (German), and Belgium (Dutch, French, German, English). Envisaged regional divisions include one for Romania, Bulgaria, Cyprus, and Greece (all official languages plus French and English); one for Nordic countries (Denmark, Sweden, Finland, Estonia, Latvia, and Lithuania, with English as the main language); and one for Hungary, Czech Republic, and Slovakia (language undecided as of January 2015).

A Drafting Committee, composed of highly experienced judges and lawyers, is in charge of preparing the final draft. As of January 2015, the 17th version has been released (up-to-date information can be found on www.unified-patent-court.org). Areas of concern include a potential transitional regime, especially regarding an opt-out (from the Unitary Patent) provision; the possibility or not to bifurcate; the working languages at various divisional levels; the training, qualifications, and experience of judges; and court fees, recoverable costs, and numerous further procedural details. If bifurcation is allowed, there will be four different ways to enforce patent protection in Europe (Hilty et al. 2012):

1. Patents granted by national patent offices and enforced through national courts
2. European Patents granted by the EPO and enforced through the Unified Patent Court system
3. European Patents granted by the EPO and enforced through national courts
4. Unitary Patents granted by the EPO and enforced through the Unified Patent Court system

It will be possible to switch from option 3 to option 2 as long as the patent has not been subject to litigation in a national court. This new enforcement system and its flexibility will pave the way for more strategic behavior by patent owners. Unitary Patents might be allowed to bifurcate toward European Patents and hence might actually not be enforced only through the UPC. This bifurcation issue is, as of January 2015, still subject to change.

Longer-Term Challenge: Make It Work. . .

The European patent system which is being envisaged by policy-makers will include three layers of patents (national, European, and Unitary) and four enforcement mechanisms. Its cost in terms of renewal fees could be particularly high, not to mention the litigation costs and the complexity of the whole system. True, the Unitary Patent Package is in itself less complex and expensive in relative terms (i.e., per capita or per market unit), but it is being built on top of a two-layer system.

One can hardly disagree with the fact that this system, in its current format, will probably not better fulfill its ultimate objective of stimulating innovation in Europe: three layers make it more complex, and it might end up being quite expensive. This three-layer system is to be compared with a less expensive one-layer system in the rest of the world. But this is certainly not a reason to discontinue the Unitary Patent and Unified Patent Court projects. They are crucial for the construction of the European patent system, and much political power or courage will be needed to

achieve a coherent and effective system. In order to build an effective and truly European patent system, the following milestone will have to be considered:

1. National patent offices should stop granting patents on their own, which does not preclude playing an important role in their national innovation system, to continue to receive national priority filings and to perform search for prior art and preliminary assessments.
2. The European Patent should be void, and a phasing-out agenda should be established.
3. Small and medium entities should have substantially lower fees and litigation costs.

References

Danguy J, van Pottelsberghe de la Potterie B (2011a) Cost-benefit analysis of the community patent. J Benefit Cost Anal (Berkeley Electronic Press) 2(2):1–41

Danguy J, van Pottelsberghe de la Potterie B (2011b) Patent fees for a sustainable EU patent system. World Patent Inf 33(3):240–247

Danguy J, van Pottelsberghe de la Potterie B (2014) The policy dilemma of the unitary patent. Bruegel working paper 2014/13, 27 Nov 2014

de Rassenfosse G, van Pottelsberghe de la Potterie B (2013) The role of fees in patent systems: theory and evidence. J Econ Surv 27(4):696–716

Guellec D, van Pottelsberghe de la Potterie B (2007) The economics of the European patent system. Oxford University Press, Oxford, 250 p

Hilty R, Jaeger T, Lamping M, Ulrich H (2012) The unitary patent package: twelve reasons for concerns. Max Planck Institute for Intellectual property and competition law research paper, 17 Oct 2012

Mejer M, van Pottelsberghe de la Potterie B (2011) Patent backlogs at USPTO and EPO: Systemic failure vs deliberate delays, World Patent Information, 33 (2):122–127

Mejer M, van Pottelsberghe de la Potterie B (2012) Economic incongruities in the European patent system. Eur J Law Econ 34(1):215–234

van Pottelsberghe de la Potterie B (2010) Europe should stop taxing innovation. Bruegel policy brief, 2010/2, 8 p

Van Pottelsberghe B, van Zeebroeck N (2008) A brief history of space and time: the scope-year index as a patent value indicator based on families and renewals. Scientometrics 75(2):319–338

European Union Anti-cartel Policy

J. M. Ordóñez-de-Haro[1], J. R. Borrell[2,3] and J. L. Jiménez[4]

[1]Departamento de Teoría e Historia Económica y Cátedra de Política de Competencia, Universidad de Málaga, Málaga, Spain

[2]Secció de Polítiques Públiques, Dep. d'Econometria, Estadística i Economia Aplicada, Grup de Governs i Mercats (GiM), Institut d'Economia Aplicada (IREA), Universitat de Barcelona, Barcelona, Spain

[3]IESE Business School, Public-Private Sector Research Center, University of Navarra, Navarra, Spain

[4]Facultad de Economía, Empresa y Turismo, Departamento de Análisis Económico Aplicado. Despacho D.2-12, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

Abstract

The fight against cartels started in European Union since its foundation in 1957 and the passing of the Regulation empowering the Commission to enforce competition rules since 1962. However, anti-cartel policy was very ineffective in its early years from 1962 to 1980. It also had a lot of enforcement problems to uncover cartels until 1995. It was in 1996, when the leniency program was set up, when it truly became an increasingly effective policy. The leniency program is a mechanism by which infringing firms that have been active in a cartel can obtain fine reductions by providing hard evidence to the Commission about the existence and functioning of any cartel. The improving of the leniency program in 2002 and 2006, and the adoption of tougher fining policies and settlement procedure, has made cartel-busting policy much more effective in the last decade, but there is still much uncertainty to what extend anti-cartel policy will keep this trend of being more effective in the future.

EU Anti-cartel policy stages

The fight against cartels started in European Union (EU) since its foundation in 1957 and the passing of the Regulation empowering the Commission to enforce competition rules since 1962. Though in the last decade a clear advance in the detection, destabilization, and fining of cartels operating in the European Union has been noticed, the first years of anti-cartel policy were flawed. In fact, as Ordóñez de Haro et al. (2016) argue, four stages in EU anti-cartel policy can be highlighted.

First Stage (1957–1980)

The origin of the EU competition policy is found in the relevant provisions of the 1957 Treaty of Rome. The Article 85 of this Treaty (now Article 101 of the Treaty on the Functioning of the European Union) bans agreements among firms that restrain competition in the internal market.

The actual enforcement of these articles 85 and 86 became effective as of 21 February 1962, with the entry into force of the Council Regulation n° 17/62. This Regulation gave the Commission a central role as the authority charged with enforcing those articles, recognizing its power to open investigations, to adopt decisions and impose appropriate sanctions and remedies for the competition rules' infringements (McGowan 2009; Carree et al. 2010).

The enforcement system was based on the Commission's centralized control of the application of Article 85 and the requirement of prior notification by the parties of their agreements, decisions, and practices to the Commission which after examination of the notification could authorize them based on the proper exemptions contained in paragraph 3 of Article 85. This system caused serious delays in the procedural treatment and completion of the files and the consequent backlog of cases, since the Commission devoted a large proportion of its resources to deal with notifications.

To address this problem, the Commission tried to reduce the number of cases and to speed up the decision-making procedure by undertaking several initiatives, including the adoption and application of several block exemption regulations, the use of the so-called comfort letters, or the introduction of notices on agreements of minor importance which do not have sufficient impact on competition.

At the end of this stage, in its decision in case *1V/29.595 – Pioneer Hi-Fi Equipment*, the Commission expressed its determination to shift its fining policy toward harsher sanctions for competition law infringements (Geradin and Henry 2005).

Seven cartels were sanctioned in this early stage of the European anti-cartel policy, totaling 55.53 million constant 2010 euro in fines (Ordóñez de Haro et al. 2016). The Commission just completed 0.4 sanctioned cases per year since 1962–1980 and imposed just fines of 600,000 euros per participating parent firm on average (constant 2010).

Second Stage (1981–1995)

It was a transitional stage during which the first consequences of that shift in Community fining policy were seen in the general level of fines imposed on cartels. The Commission reiterated its intention to move towards a tougher sanctioning policy in its Thirteenth Report on Competition Policy (1984).

Although there were no relevant legislative or institutional developments during this period, in December 1995, the European Commission published a draft notice concerning the non-imposition or mitigation of fines in cartel cases where undertakings cooperate in the preliminary investigation or proceedings in respect to an infringement. The subsequent adoption of this notice in 1996 represents a major milestone in EU competition law enforcement against cartels.

Over these 15 years, the Commission sanctioned 32 cartel cases. The total amount of fines exceeded 1122 million euro (constant 2010) in this period. The Commission just completed 2.1 sanctioned cases per year in this period and imposed fines of 4 million euro per participating parent firm on average (constant 2010).

Third Stage (1996–2005)

The introduction of the first leniency program in 1996 marks the start of a third stage which covers

all the steps taken toward the modernization of the EU competition policy up to 2005.

The first Community leniency program started on 18 July 1996. It was inspired by the US program in force since 1993, sought to encourage the breakdown of the "code of silence" among the members of the cartels and to become a successful tool that would significantly increase the effectiveness of competition policy (see Borrell et al. 2014).

Although this first leniency system is characterized by introducing some of the basic principles that would guide subsequent versions of the program, it lacked many enforcement details that were fixed in the 2002 and 2006 reforms of the leniency program.

Another significant change in this period was the publication by the Commission in January 1998 of its "*Guidelines on the method of setting fines imposed pursuant to Article 15 (2) of Regulation No 17 and Article 65 (5) of the ECSC Treaty*." The European Commission, for the first time, provided a clear method for determining the final amount of a fine explaining the successive steps and criteria that would be taken in order to obtain that amount.

The creation of the first "anti-cartel unit" within DG IV (directorate later known as DG Competition), in December 1998, represented an important reinforcement of the particular fight against cartels. The progressive increase in resources culminated in the creation of a second "anti-cartel unit" in 2002. Nevertheless, the two anti-cartel units ceased to exist as they were conceived as a result of an internal reorganization of the DG Competition in July 2003. From then on, the resources and staff to combat cartels were then allocated to different units in several key sector directorates (Lowe 2008).

The numerous criticisms of the manifest deficiencies in the *1996 Leniency Notice* yield to the publication, on 19 February 2002, of the new *2002 Leniency Notice* which brought about significant changes to the procedure and specified requirements to benefit from the program (Arp and Swaak 2003; Borrell et al. 2015).

During the period covered by this stage, the provisions contained in the Council Regulation n° 17/1962 were valid until May 2004 when the *Council Regulation (EC) No 1/2003 of 16 December 2002 on the implementation of the rules on competition laid down in Articles 81 and 82 of the Treaty* entered into force. The most significant changes introduced by the Regulation 1/2003 mean the simplification of the administrative procedures and decentralization of the application of the competition rules in the EU. This reform implies that the system of notification and authorization was completely abandoned to be replaced by a directly applicable exception system. The Commission no longer had to deal with notifications from firms taking part in agreements, but the firms were directly liable of their conducts before the Commission, the National Antitrust Authorities, the European Courts, and the National Courts if they infringe the provisions of the Treaty reducing the Commission's workload significantly (McGowan 2005).

From 1 June 2005 on, in response to the greater emphasis on combating cartels, a new Directorate in the DG Competition devoted exclusively to the fight against cartels became operational. The Cartels Directorate is responsible for carrying most of the cartel cases and plays a leading role, in close cooperation with the Directorate for Policy and Strategic Support in developing the policy to apply in the cartels arena.

Additionally, in 2005, a two-stage procedure is introduced by the DG Competition. So, from then on, all antitrust cases start with a first phase of investigation after which the Commission adopts a decision concerning the theory of identified harm and whether there is reason to pursue the case as a matter of priority. If the case is considered a priority, the Commission takes a decision in principle to initiate proceedings and then carries out a thorough investigation. This procedure is intended to spend less time and resources on cases that do not merit special attention.

To conclude this third stage, we should note the first initiative which tried to identify the main obstacles faced by the operation of a more efficient system of claims for damages which resulted from breaches of European Union competition laws: on 19 December 2005, the Commission

published the *Green Paper – Damages actions for breach of the EC antitrust rules* which includes the proposal of a series of measures aimed at encouraging victims of infringements of Articles 81 and 82 of the Treaty to exercise their right to claim for damages before Member States' courts (Pheasant 2006). Consequently, this initiative is designed to boost the interaction between public and private enforcement of antitrust rules (Diemer 2006).

In this period, a total of 44 cartel cases were sanctioned, 46% of them uncovered by a leniency application. Total fines imposed reach 5491.74 million constant euro. The Commission completed 4.4 sanctioned cases per year in this period (more than doubling the number of the previous stage) and imposed fines of 22 million constant euro per participating parent firm on average, multiplying the average fine by more than five times with respect the previous period.

Fourth Stage (2006–)

The last stage began in 2006, with several legislative reforms that aim to improve and consolidate those tools which proved to be valuable for fighting cartels, such as the sanctioning mechanisms, as well as the framework for cooperation between cartel participants and the Commission, in order to streamline procedures and to enhance the transparency and predictability.

The first of these improvements took place with the publication, on 1 September 2006, of the *2006 Fine Setting Guidelines*, which the Commission is currently putting in place, and they introduce some important new points (Barbier de La Serre and Lagathu 2013).

On 8 December 2006, the *2002 Leniency Notice* was replaced by the new *2006 Leniency Notice* (OJ C 298/17. 8.12.2006.). This is the second reform of the leniency program and aimed to provide for more clarity and transparency in its requirements and how to proceed in this program, as well as make it more attractive to potential cooperators (Sandhu 2007; Borrell et al. 2015).

In early July 2008, the Commission adopted the so-called Settlements Package. It consists of two documents: the *Commission Regulation (EC) No 622/2008 of 30 June 2008 amending Regulation (EC) No 773/2004*, as regards the conduct of settlement procedures in cartel cases and the *Commission Notice on the conduct of settlement procedures in view of the adoption of Decisions pursuant to Article 7 and Article 23 of Council Regulation (EC) No 1/2003 in cartel cases*.

The *Settlement Notice* details the settlement procedure, interprets the new provisions contained in the *Settlement Regulation*, and provides further guidance on the first steps of the settlement procedure, the settlement discussions, the statement of objections, and the adoption of the final decision in such a procedure.

This package enables the Commission and parties to proceedings to follow a more simplified and simpler procedure when cartel participants, having seen the evidence in the Commission file, acknowledge their involvement in the cartel and their liability for the infringement (Mehta and Tierno 2008).

Finally, it is essential to stress the possible consequences for the fight against cartels that the adoption of the *Damages Directive* has and could have. Member States have until 27 December 2016 to transpose the provisions of the Directive into their legal systems (*Directive 2014/104/EU of the European Parliament and the Council of 26 November 2014 on certain rules governing actions for damages under national law for infringements of the competition law provisions of the Member States and of the European Union*). It remains to be seen whether the setup of an EU wide common framework for claiming private damages will make this policy even more effective in the near future.

The Commission's sanctioning activity during this stage was the most prolific of all with 53 cartel decisions until the end of 2014. Total fines imposed reach 17,739.87 million euro constant 2010. The Commission completed 5.9 sanctioned cases per year in this period and imposed fines of 62 million euro per participating parent firm on average (constant 2010), almost multiplying the average fine by three times with respect the previous period.

References

Arp DJ, Swaak CR (2003) A tempting offer: immunity from fines for cartel conduct under the European Commission's new leniency notice. Eur Compet Law Rev 24(1):9–18

Barbier de La Serre E, Lagathu E (2013) The law on fines imposed in EU competition proceedings: faster, higher, harsher. J Eur Compet Law Pract 4(4): 325–344

Borrell JR, Jiménez JL, García C (2014) Evaluating antitrust leniency programs. J Compet Law Econ 10(1):107–136

Borrell JR, Jiménez JL, Ordóñez de Haro JM (2015) The leniency programme: obstacles on the way to collude. J Antitrust Enfor 3(1):149–172

Carree M, Günster A, Schinkel MP (2010) European antitrust policy 1957–2004: an analysis of commission decisions. Rev Ind Organ 36:97–131

Diemer C (2006) The green paper on damages actions for breach of the EC antitrust rules. Eur Compet Law Rev 27(3):309–316

Geradin D, Henry D (2005) The EC fining policy for violations of competition law: an empirical review of the commission decisional practice and the community courts' judgments. Eur Compet J 1(2): 401–473

Lowe P (2008) The design of competition policy institutions for the 21st century-the experience of the European Commission and DG Competition. Compet Policy Newsl 2008-3:1–11

McGowan L (2005) Europeanization unleashed and rebounding: assessing the modernization of EU cartel policy. J Eur Publ Policy 12(6):986–1004

McGowan L (2009) Any nearer to victory in the 50-year war? Assessing the European Commission's leadership, weapons and strategies towards combating cartels. Perspect Eur Polit Soc 10(3): 283–307

Mehta K, Tierno ML (2008) Settlement procedure in EU cartel cases. Compet Law Int 4:11–16

Ordóñez de Haro JM, Borrell JR, Jiménez JL (2016) European Commission's fight against cartels (1962–2014): a retrospective and forensic analysis. Mimeo

Pheasant J (2006) Damages actions for breach of the EC antitrust rules: the European Commission's green paper. Eur Compet Law Rev 27(7):365–381

Sandhu JS (2007) The European Commission's leniency policy: a success? Eur Compet Law Rev 28(3): 148–157

European Union Law

► European Community Law

Europeanization

► European Nationality

Experimental Law and Economics

Sven Hoeppner
Center for Advanced Studies in Law and Economics (CASLE), Ghent University Law School, Ghent, Belgium
Guest Researcher, Max Planck Institute for Research on Collective Goods, Bonn, Germany

Abstract

Experimental law and economics is the newest methodological development in law and economics research. Yet a lot of researchers – old and young – are not familiar with the foundations of the method. This may lead to skepticism. In this entry, I elaborate on the purpose of experiments and introduce building blocks of the method that tend to distinguish experimental law and economics from experimental methods in other disciplines. Moreover, I shortly discuss concerns about the external validity of the results obtained in a laboratory experiment. Finally, I introduce and invalidate the most common criticisms against experiments that most often advanced by scholars unfamiliar with the method and the discipline. As knowledge about the method's foundations is further spread, experimental law and economics will substantially contribute advancing the research frontier of the economic theory of law.

Definition

Experimental law and economics exposes theoretical research in law and in law and economics to methods from experimental economics, thereby testing existing theory, investigating observed

deviations from theory, establishing empirical regularities as the foundation of new theoretical constructs, and informing policy-making.

Introduction

Not every lawyer, legal policy-maker, judge, or legal scholar conceives legal rules as a governance instrument that can be employed to coordinate and steer individual behavior in society. For those who do adopt such an ex ante perspective, however, it is of utmost importance to hypothesize about and investigate how law and individual human behavior are interrelated. To pierce the veil of ambivalent plausibility arguments and conflicting theories, it is even more important to understand what kind of behavior may be caused by alternative legal interventions.

Yet, legal scholars are often surprised and become quite reserved when they learn that a colleague is not extensively engaged in doctrinal legal scholarship, but rather adopts a social science approach. Caution usually rises to suspicion when the straying colleague conducts experiments, which seem rather alien to legal scholarship. But experiments can, inter alia, help to understand whether policy proposals – developed through, e.g., doctrinal legal scholarship – will function as intended.

Actually, experiments are not a novel method in legal scholarship. They have long been conducted in the law and psychology as well as in the criminology branches of legal research (Engel 2013a, b). Today, "[t]he only true novelty is the advent of experimental law and economics" (Engel 2013a, p. 7).

This entry is not written as a review of studies using an experimental approach to law and economics. Excellent reviews with a cornucopia of examples exist (cf. Croson 2002; Arlen and Talley 2008; Lawless et al. 2010). Rather, it is written with the intention to confront all too common misunderstanding about the experimental approach to law and to law and economics with information about the usefulness of experimental methodology for legal scholarship as well as policy-making. This entry should provide a good first reading for an audience interested – but yet unfamiliar – with experimental approaches to law (and economics).

Purpose

Generally speaking, in both disciplines, economics and law, there is a strong emphasis on deductive reasoning. Researchers start with a set of axioms, use rules of logic to manipulate them, and subsequently derive conclusions about their research question. For instance, in (neoclassical) economics, certain assumptions about individual behavior are used to mathematically model the theoretical construct "market" and to draw implications for the allocation of resources. As an example for deductive reasoning in law, one core assumption is that the law makes sense (ratio legis) and a lot of doctrinal legal scholarship is devoted to establish this inner logic of the law by means of interpretation and, furthermore, to match real-world facts to these abstract norms by analytical comparison (subsumption). Thus, doctrinal legal scholarship can conclude that legal consequences from an abstract, model-like norm are applicable to real-world facts.

A consequence of this heavyweight deductive reasoning is that theories become so complex and refined that simply observing natural data does often not address the pressing theoretical questions. Empirical tests of those theories are therefore significantly more difficult to perform. Experiments are one way to meet such a challenge (Croson 2002).

The more conducive purpose of conducting experiments in law, in general, and law and economics, in particular, is to exploit three advantages of the methodology. The first, and main advantage of experiments, lies in their possibility to solve the identification problem – the problem that arises when there may be bidirectional influence between dependent and independent variables that makes it difficult to separate cause and effect. For example, consider crime levels and the number of police: Does an increase in the level of crime cause a demand for more police, or do

fewer police cause more crime? Identification is notoriously difficult with field data. To establish causality outside of an experiment, at least three steps are necessary (Lawless et al. 2010). First, there needs to be a reliable association – or strong correlation – between the two variables. Second, one needs to credibly show that changes in the variable hypothesized to be the cause precede the supposed effect variable. Third and most problematic, other rival explanations for this observation (confounds) need to be ruled out, which also concerns the rather big problem of omitted-variable bias, where a driving variable is missing from the statistical model used for analysis. As a consequence, the effect of variables that enter the analysis is over- or underestimated.

To the end of separating cause and effect, a laboratory experiment generates an artificial environment in which participants are randomly assigned to different treatments. Random assignment facilitates control for unobservable individual differences among the participants. If carefully designed, the treatments in the experiment differ only by one dimension. Accordingly, if a treatment effect can be observed, it is a necessary result from the manipulation. With their distinguishing features – manipulation of one or more variables of interest, random assignment, and experimental control of other variables – experiments are specifically designed to test cause-effect relationships (Lawless et al. 2010). The very large degree of control can be used to generate a laboratory environment that tests theories and separates alternate theories that may not be separable with naturally occurring data (cf. Croson 2002).

The second main upside of laboratory experiments is their replicability. Because laboratory experiments are replicable, other researchers can reproduce the experiment and verify the findings independently. A stream of independent replications can test the robustness of findings. This is very difficult with field data because the environment is constantly changing such that each observation or dataset is just a snapshot of a certain combination of conditions.

Especially for legal scholars, experimentation brings about a third important benefit that is likely to be equally important: experiments enable the study of problems that are hard or even impossible to observe in the field (Lawless et al. 2010), i.e., even if data are already available from other sources. How would addressees of a legal rule act if the existing rule would be or would have been replaced by an alternative rule? And what are the behavioral effects of different alternative rules that are under consideration for being implemented? Many a fact that is relevant for doctrinal lawyers and legal policy-makers are counterfactuals and never enter into field data (Engel 2013a). Experimenters can observe the unobservable (Falk and Heckman 2009).

Building Blocks

The basic logic of an experiment is to create different treatment groups (or experimental conditions) by manipulating one or more variables of interest, but always only one at a time, while every other variable is experimentally controlled for. The treatment without any variable change serves as the control group. This experimental manipulation creates desired differences between the groups, which enables the isolation of the different variables of interest, reduces confound, and ultimately facilitates identification of cause-and-effect relations. Participants are randomly assigned to these treatment groups. Randomization may occur for a lot of aspects in a specific experimental design. With regard to randomly assigning participants to experimental condition, random assignment, e.g., assures the initial comparability of the different experimental groups by equalizing the distribution of non-manipulated participant characteristics, such as age, income, gender, and education, across the experimental conditions. While participants are then being exposed to the experimental stimuli, their decision behavior is measured. Afterward, the different measurements can be compared across treatment groups. Because the only difference between the groups is their exposure to the treatment, any measured difference between the treatment and control groups can be attributed to the treatment.

Beyond this general categorization, experimental methodologies and their best practices – even within the social sciences – are quite different. Because of different underlying theories, experiments in a law and economics paradigm differ in their methodology from the earlier experimental approaches to law in psychology and criminology. Experimental law and economics is based on the methodology of experimental economics (cf. Hoffman and Spitzer 1985; Davis and Holt 1993; Friedman and Sunder 1994; Kagel and Roth 1995).

Although experimental economics pushes more and more to the field (Levitt and List 2007, 2009), the standard tool in experimental economics is the laboratory experiment. The lab offers the possibility to effectively employ controlled variation at relatively low cost (Hoffman and Spitzer 1985; Friedman and Sunder 1994). Decision environments can be controlled in ways that are very difficult to duplicate in the field. In the lab, the experimenter knows and controls participants' material payoffs, the order in which the different participants can act and interact, and – maybe most importantly – the informational situation for each single choice.

In experimental economics, one can generally differentiate four types of experiments for specific purposes: some experiments test theory, others investigate observed deviations from theory (anomalies), again others establish empirical regularities as the foundation of new theories, and finally experiments inform policy-making or testbed new policy proposals (cf. Hoffman and Spitzer 1985; Roth 1986; Smith 1994). Each of these types faces particular methodological challenges (cf. Croson 2002).

The decisive difference between experimental economics – and by extension experimental law and economics – and the other experimental approaches to law involves a number of dimensions (cf. Davis and Holt 1993; Friedman and Sunder 1994; Engel 2013b).

1. Participants are exposed to a specific incentive structure that matches the payoffs in the theory underlying the experiment. The choices the participants make in the experiment determine their compensation. This ensures that participants seriously contemplate their decision because value is induced (Smith 1976; Friedman and Sunder 1994). This is the main reason why experimenters in economics trust that "behavior in the laboratory is reliable and real: Participants in the lab are human beings who perceive their behavior as relevant, experience real emotions, and take decisions with real economic consequences" (Falk and Heckman 2009, p. 536).
2. Since real money is at stake, participants do not face a hypothetical, but rather a real situation. Note that in some cases and contradictory to established experimental economics, it may make a lot of sense to not incentivize a participant. This is the case, for instance, if the participant mimics the role of a judge who has, ideally, no monetary incentive in the single decision in real life (cf. Engel and Kurschilgen 2011).
3. The experiment is usually free from context. Economic experiments are very abstract to facilitate identification (cf. Ariely and Norton 2007). Firstly, context may add variance to the data and blur otherwise clear results. More worrisome, context may create systematically biased results or demand effects, i.e., when the participant tries to act according to the perceived expectations of the experimenter. As a result, the generated dataset would not be reliable. Lastly, avoiding context is, of course, a requisite when the underlying theories are also free from contextual cues. Note that context, however, may be a treatment variable as well. If so, the manipulation of context is the premier goal of the experiment.
4. Another distinguishing feature is that experimental economists generally take great care to not deceive their subjects. This is different in psychology experiments where deception is commonplace, although not undisputed (cf. Hertwig and Ortmann 2008). In experimental economics, the validity of the results depends on the link between expected payoffs and individual behavior. If participants are deceived about variables establishing this link, the validity of their decisions is

questionable. Moreover, experimental economists usually avoid tainting the participant pool. They are "concerned about developing and maintaining a reputation among the student population for honesty in order to ensure that subject actions are motivated by the induced monetary rewards rather than by psychological reactions to suspected manipulation" (Davis and Holt 1993, pp. 23–24). Admittedly, however, there is an intense, contemporary discussion about whether the costs of deception in experimental economics exceed the benefits (cf. Bonetti 1998; Hertwig and Ortmann 2008; Jamison et al. 2008; Alberti and Güth 2013; Krawczyk 2013).

5. Economic experiments are oftentimes interactive in nature. The reason for interaction between participants is straightforward. Economics experiments mostly employ some form of social dilemma game as experimental workhorse, and theoretical predictions focus on the results of social interaction.
6. Usually, experimenters in economics take great effort to ensure complete anonymity. As with all the other features, this is done with the aim to facilitate identification, i.e., to avoid confounds. To provide a counterexample, the experimentalist may not care about absolute anonymity, if the research question addresses reputation effects.
7. Often, the trials in the experiment are repeated to allow for learning and to study interactive dynamics.

These features facilitate what an experiment is ultimately all about: complete control. Only in an environment with extensive control can the claim become credible that an observed difference in behavior is indeed caused by the treatment manipulation and not merely the result of correlation.

Caution

However promising and rewarding the clear separation of cause and effect, experimenters face validity concerns. In general, the concept of validity refers to the extent to which an empirical study produces accurate and credible data. Validity is assessed on three dimensions. First, internal validity refers to the extent to which the research design allows making inferences about the relationship of different variables. Internal validity involves that the experimenter must ensure that the choices faced by the control and treatment groups differ in the ways the experimenter hypothesizes, but do not differ in some other ways. This implies accounting for alternative explanations and omitted variables. Second, external validity speaks to the degree to which the research results can be generalized to a population beyond the particular study, i.e., to different people, different settings, different times, and also different measures. Third, construct validity refers to the extent to which the measures used to observe certain variables sufficiently capture the construct that the empiricist wants to study (cf. Lawless et al. 2010).

In order to isolate variables and to identify cause-effect relationships, experimentalists necessarily trade off external validity for internal validity. The artificial environment in a laboratory experiment allows the experimenter to make clean-cut inferences, but raises questions about its congruence with the social phenomenon that is to be studied. By design, what is measured in a laboratory experiment is only analogous to what the experimenter wants to understand (Smith 2010; Engel 2013b). Just because a hypothesis is falsified in the lab does not necessarily suggest that it will also be falsified in the field because there are so many intervening variables that cannot be controlled for. However, Plott (1982) counters concerns about external validity by pointing out that a theory will likely not predict outcomes in the real world if it does not (at least) predict outcomes in an idealized, controlled laboratory environment. Falk and Heckman (2009) argue that external validity problems are not a problem unique to experimental methodology.

Especially when legal scholars are on the hunt for cause-and-effect relationships between legal rules and individual behavior, they pay an increased price because they tend to distance themselves even more from the traditional doctrinal legal discourse (Engel 2013a).

Criticism and Concern

Particularly when scholars or members of another audience are unfamiliar with experimental work in general and the specific methodology in particular, a standard set of criticisms is often held against the experimental results.

Probably the most notorious objection involves limited external validity (see above). The gist of this critique is that the experiment cannot tell anything about the real world because the decisions are made in an artificial environment. As mentioned above, this is indeed reason to be very cautious to interpret experimental findings. However, this typical reason to disapprove of experimental results is misguided if the experiment satisfies the assumptions of the tested theory. Theoretical predictions rest on the assumptions of that theory irrespective of where the assumptions are met – in the lab or in the field (Plott 1982). The objection about limited external validity may have more grips, when the experiment concerns legal policy-making. In this context, an isolated cause-effect relationship matters less than entire institutions where certainly many of those relationships work with or against each other. Nevertheless, this implies the need to understand complex institutions by disentangling the muddle of causes and effects. Alternatively, legal experimenters may implement sequential designs that investigate behavior before and after the implementation of the entire legal institution (Engel 2013b). Lastly, empirical scholars know that likely none of the available methods will ever fully address reality. Falk and Heckman (2009) emphasize that it is not so obvious whether laboratory data or the field data are more informative. The scientific value of a method ultimately depends on the underlying research question. The empirical methodology notwithstanding, "eventually, reality is too complex, too little orderly, to be studied in an objective way" (Engel 2013a, p. 18). The concern about external validity certainly puts another emphasis on the complementary nature of different empirical methods, but it provides no grounds to completely dismiss one of them.

A second standard objection against experimental results concerns the participant pool. Novices in experimental methodology often criticize the fact that experiments in law and economics more often than not rely on university students. They argue that students are certainly very different from the population in the real world and that they make decisions differently from professional decision-makers outside the lab. Fortunately, experimental research on this subject is on the rise. Oftentimes, there is no discernible difference in decision-making between students and professionals. If there is, however, students sometimes are more successful than professionals – although not always. There are a number of reasons why this objection seems to have a weak foundation (cf. Croson 2002; Falk and Heckman 2009). A student sample is as representative as any other sample, as long as it does not consists of only students with the same academic background, if this is part of the manipulation (cf. Frank et al. 1993, 1996). Most importantly, most often, the experiment is not reliant on any specific knowledge or experience. By contrast, there are also risks in bringing professionals to the lab. For instance, they may bring in incentives, knowledge, and experience that do not exist in the experimental design and therefore systematically bias results. Nevertheless, it is surprising that people often do not complain about seemingly weak participant pools in other areas of science. This is nicely illustrated with a not so ironic insight from Gary E. Bolton that he shared with me and others in July 2013: "In experimental economics, we typically begin our studies with students, pretty much for the same reasons medical experimenters begin with mice. First, they work cheap. Second, we can do what we want with them (within institutional rules). And third, they are genetically similar to real human beings".

Another prominent critique of experimental work in law and economics is based on the size of the payoffs in the experiment. It is often argued that the size of payoffs is too small to induce meaningful behavior. However, experiments have been conducted with low and high payoffs and, in fact, have even tested differences between low and high stakes in experimental settings. All in all, there are very few differences on average behavior when the size of the stakes varies (Smith

and Walker 1993; Beattie and Loomes 1997; Camerer and Hogarth 1999).

Taken together, these common objections can be summarized as the "lack of realism" challenge. This notion, however, is based on an implicit assumption about the hierarchy of how relevant data are generated, with field data being deemed superior to data from the laboratory. Ultimately, however, the issue of realism is not a question of laboratory versus field data. The real issue is determining the best way to identify the causal effects in question. In this context, it is also important to stress that empirical methods are complements to each other (cf. Arlen and Talley 2008), not substitutes, as the discussion about the "lack of realism" suggests. Paradoxically, however, many of these objections also suggest conducting more experiments instead of fewer (Falk and Heckman 2009), because the critique itself is often based on a variable that can be manipulated in a new set of experiments to test the critique. This is, for instance, true for the aforementioned criticism that stakes are trivial and that participants are inexperienced students. However, this is not to suggest that experimentalists should not take these common concerns seriously. On the contrary, experimentalists should be their own most austere critics in order to prevent automatic, arbitrary experimental design that would threaten the credibility of experimental results.

In addition to these standard objections to the experimental method, Engel (2013a) specifically discusses sets of concerns prevalent in the legal academic arena. He distinguishes the rather diffuse concerns into five common categories of specific objections from legal scholars. To traditional lawyers, the experimental method appears to be (1) too scientific, (2) too individualistic, (3) too narrow, (4) too anxious, and (5) too small. While cleanly dissecting and acknowledging the foundations of these concerns, Engel (2013a) humbly refrains from completely resolving – or, for that matter, rejecting – them. Also, doctrinal legal scholarship and experimental law and economics can achieve more, if they work hand in hand.

Summary

This entry has introduced the purpose of conducting experiments – also in a legal context – and the basic building blocks of the experimental law and economics paradigm. A cautionary emphasis has been put on the validity trade-off that is inherent in all experimentation. Finally, criticism and objections have been briefly surveyed. These need to be taken – objectively – into account, if experimentalists care for the scientific value of their own method.

As legal scholars, "[w]e will quite likely have to live with some tension between science and doctrine, between individualistic and social constructions of legal problems, between the conditions for causal inference and lumpy institutional choice, between the responsibility for legal development and the competitive pressure of peer-review, between intuition and explication" (Engel 2013a, p. 29). Yet, experimenters can bring to light scientific knowledge that matters for the adoption of new rules, the further development of existing rules, and even the decision of concrete cases. Therefore, doctrinal legal scholars, lawyers, judges, and regulators should find it at least useful and relevant to have the generic knowledge available that has been generated in the laboratory.

In offering a mere introduction and a tertiary source, this entry necessarily stayed at the surface of what can only be called a fascinating methodological challenge. If this text kindled an initial spark among the audience, it was utterly successful. As is the case for the other empirical research methods, however, experiments are not to replace traditional legal discourse. Rather, they are promising and enlightening complements to doctrinal legal scholarship and the continuous evolution of the law and its economic theory.

Cross-References

- ▶ Empirical Analysis
- ▶ Games
- ▶ Good Faith and Game Theory
- ▶ Knowledge
- ▶ Panel Data Analysis

References

Alberti F, Güth W (2013) Studying deception without deceiving participants: an experiment of deception experiments. J Econ Behav Organ 93:196–204

Ariely D, Norton MI (2007) Psychology and experimental economics: a gap in abstraction. Curr Dir Psychol Sci 16:336–339

Arlen JH, Talley EL (2008) Experimental law and economics. Edward Elgar, Cheltenham

Beattie J, Loomes G (1997) The impact of incentives upon risky choice experiments. J Risk Uncertain 14:155–168

Bonetti S (1998) Experimental economics and deception. J Econ Psychol 19:377–395

Camerer CF, Hogarth RM (1999) The effects of financial incentives in experiments: a review and capital-labor-production framework. J Risk Uncertain 19:7–42

Croson R (2002) Why and how to experiment: methodologies from experimental economics. Univ Ill Law Rev 2002:921–945

Davis DD, Holt CA (1993) Experimental economics. Princeton University Press, Princeton

Engel C (2013a) Legal experiments: mission impossible? Eleven International, The Hague

Engel C (2013b) Behavioral law and economics: empirical methods. Preprints of the Max Planck Institute for Research on Collective Goods Bonn, 2013/1

Engel C, Kurschilgen M (2011) Fairness ex ante and ex post: experimentally testing ex post judicial intervention into blockbuster deals. J Empir Leg Stud 8:682–708

Falk A, Heckman JJ (2009) Lab experiments are a major source of knowledge in the social sciences. Science 326:535–538

Frank RH, Gilovich T, Regan D (1993) Does studying economics inhibit cooperation? J Econ Perspect 7:159–171

Frank RH, Gilovich T, Regan D (1996) Do economists make bad citizens? J Econ Perspect 10:187–192

Friedman D, Sunder S (1994) Experimental methods: a primer for economists. Cambridge University Press, Cambridge

Hertwig R, Ortmann A (2008) Deception in experiments: revisiting the arguments in its defense. Ethics Behav 18:59–92

Hoffman E, Spitzer ML (1985) Experimental law and economics: an introduction. C Law Rev 85:991–1036

Jamison J, Karlan D, Schechter L (2008) To deceive or not to deceive: the effect of deception on behavior in future laboratory experiments. J Econ Behav Organ 68:477–488

Kagel JH, Roth AE (1995) The handbook of experimental economics. Princeton University Press, Princeton

Krawczyk M (2013) Delineating deception in experimental economics: researchers' and subjects' views. University of Warsaw Faculty of Economic Science Working Papers, no 11/2013 (96). Online available at: http://www.wne.uw.edu.pl/inf/wyd/WP/WNE_WP96.pdf. Retrieved 05 June 2014

Lawless RM, Robbennolt JK, Ulen TS (2010) Empirical methods in law. Aspen, New York

Levitt SD, List JA (2007) What do laboratory experiments measuring social preferences reveal about the real world? J Econ Perspect 21:153–174

Levitt SD, List JA (2009) Field experiments in economics: the past, the present, and the future. Eur Econ Rev 53:1–18

Plott CR (1982) Industrial organization theory and experimental economics. J Econ Lit 20:1485–1527

Roth AE (1986) Laboratory experimentation and economics. Econ Philos 2:245–273

Smith VL (1976) Experimental economics: induced value theory. Am Econ Rev 66:274–279

Smith VL (1994) Economics in the laboratory. J Econ Perspect 8:113–131

Smith VL (2010) Theory and experiment: what are the questions? J Econ Behav Organ 73:3–15

Smith VL, Walker JM (1993) Monetary rewards and decision costs in experimental economics. Econ Inq 31:245–261

Exploring the Deterrent Impact of Financial Supervisory Liability

Robert J. Dijkstra
Tilburg Institute for Private Law, Tilburg University, Tilburg, The Netherlands

Abstract

This entry provides a law and economics analysis of financial supervisory liability. It discusses the deterrent impact of financial supervisory liability by using existing law and economics theory and empirical evidence.

JEL Classification

K13

Introduction

Financial supervisory authorities have the complex task of supervising the financial markets to safeguard the stability of the system as well as to ensure an effective consumer protection. Their main activities consist of licensing of financial institutions (safeguarding entry to the market),

the ongoing monitoring of the health of financial institutions, sanctioning in case of noncompliance with law and regulations, crisis management, and conduct of business supervision (Lastra 2006). When they fail in their activities, they run the risk of being held liable by both third parties and the financial institutions subject to their supervision (Tison 2003; Athanassiou 2011).

In general, public authorities are liable in the same way as any private individual or company. This position follows from the historical fact that the Diceyan conception of the rule of law did not distinguish between public and private. Where a claimant wishes to sue another party, it should, in theory, not matter whether the defendant is a public authority or a private individual or company. It is therefore interesting to notice that nowadays more than 60% of the member states of the European Union have limited the liability of their financial supervisory authorities (Dijkstra 2012).

Why should financial supervisory authorities not be treated in the same way as any other defendant? What, if anything, justifies such a special treatment? The main policy argument, commonly used in favor of shielding financial supervisory authorities from liability is the "defensive conduct" argument. This argument assumes that the imposition of liability will inhibit the effective operations of a financial supervisory authority. The fear of being held liable is so severe that authorities start to act with too much caution when dealing with the supervisee.

On the other hand, submitting financial supervisory authorities to normal liability rules can be based on the preventive function of tort law, in which the threat of being held liable gives financial supervisory authorities incentives to perform their tasks with care (Dijkstra 2009). In this way, financial supervisory liability improves financial supervision by deterring the careless execution of financial supervisory tasks.

Both sides of the debate over the impact of financial supervisory liability can be examined using a law and economics approach. This entry explores the insights that existing law and economics theory and empirical evidence offer about the most likely deterrent impact of financial supervisory liability.

The remainder of the entry is structured as follows. Section "Can We Apply Economic Analysis to Financial Supervisory Authorities?" discusses the economic analysis of public authority liability. It makes clear to what extent traditional economic analysis can be applied to public authorities and, more specifically, to financial supervisory authorities. Next, section "To What Extent Does Financial Supervisory Liability Deter?" explores the impact of liability on the behavior of financial supervisory authorities by considering the context in which these authorities operate. It describes specific factors that are likely to influence the deterrent level of financial supervisory liability. As the actual deterrent impact of financial supervisory liability is at heart an empirical question, section "Empirical Evidence" presents an overview of the existing empirical evidence. Section "Conclusion" presents the conclusion of this entry.

Can We Apply Economic Analysis to Financial Supervisory Authorities?

The prediction of the effect of a liability rule is sensitive to assumptions about the underlying behavioral model (Spitzer 1977). The party subject to liability needs to be responsive to liability claims for liability claims to have an effect on their behavior at all. Traditionally, law and economics assumes that parties behave rational and want to maximize their utility. By internalizing the costs of their actions via liability rules, they will be motivated to take optimal care and/or engage in less dangerous activities, to prevent damage from happening (Faure 2008). This model seems to fit well to private companies. These companies are generally incentivized to produce at minimum average costs in order to maximize their profits. Faced with liability, they will take preventive measures in order to avoid being held liable and experience a decrease of profits. But what about public authorities? To what extent can the traditional law and economics model also be applied to public authorities? Their objective, unlike those of private companies, is not profit maximization. Legal literature is divided when it comes to answering this question.

Some scholars have argued that public authorities face budget constraints that could result in a more or less similar response to liability as private companies (e.g., Niskanen 1971; Rosenthal 2006). This perspective can be considered as an application of the rational choice model of bureaucratic behavior which assumes that a bureaucrat wants to maximize the size of her agency's budget. Maximizing the agency's budget would provide bureaucrats with other perks they value, such as, for example, prestige, compensation, and career prospects. It has therefore been suggested that public authorities may respond to liability with behavior approximating cost minimization (Kramer and Sykes 1987).

Other scholars are more critical regarding the application of the traditional law and economics model to public authorities. Rosenthal (2006) and Levinson (2000) argue that public authorities will respond to political rather than market incentives. When the political costs of diverting public resources to loss prevention are too high, the government will not make the investment even if it is economically justified (Rosenthal 2006). According to these scholars, liability cannot be expected to promote efficient governmental investment in loss prevention. One could however argue that liability imposes, at least, some political costs. After all, tort liability diverts public funds from other activities that are politically more profitable in terms of public perception and voter's appeal. Public authority liability may, at least potentially, thus create some political incentives to prevent careless behavior. We should however take into consideration that political incentives that may be powerful at the level of elected politicians and officials might not be that powerful for the nonelected bureaucrats employed in public authorities (Dari-Mattiacci et al. 2010).

The idea that liability results in some political incentives might not be completely true when it comes to financial supervisory authorities. While most public authorities are funded from a central, governmental budget, financial supervisory authorities are, more and more, funded from fees imposed on the supervised institutions. This means that any increase of supervisory costs due to law cases against financial supervisory authorities is ultimately not fully born by the taxpayer but more and more by the financial industry, thereby diluting the impact of political incentives. One could thus question the ability of liability to motivate behavioral changes given the financial authorities' ability to pass (all or at least a significant part of) their costs on to the financial institutions subject to their supervision.

In addition to market and political incentives, one might also want to consider the effect of reputational deterrence (Gold 2016). Reputational deterrence occurs because employees want to protect their organization's pride as a form of protecting self-identity. They thus avoid actions that could expose their organization to lawsuits that can negatively affect their organizational pride. The question remains whether it is the lawsuit or the incident itself that generates negative publicity and hence negatively affects reputation.

There is obviously no easy answer to the question whether and how financial supervisory authorities respond to liability claims. Financial supervisory authorities are probably not solely self-interested or only motivated by reputational incentives, political incentives, or incentives provided by liability (Dijkstra 2010). Their behavior is likely to be influenced by their employees' own morality, social norms, organizational culture, and much more. So, at best, the behavioral consequences of imposing damages liability on public authorities are uncertain. However, the fact that more governments limit the liability of their financial supervisory authorities based on behavioral consequences makes it at least appropriate to investigate the deterrent impact of financial supervisory liability.

To What Extent Does Financial Supervisory Liability Deter?

Conditions for Deterrence

For liability to effectively deter unlawful conduct, a number of conditions have to be met. First, the targets of liability rules must be aware of and understand the rules. Secondly, they need to have the willingness to follow the rules. This will, among other factors, depend on the deterrent

level of that liability claim. In general, the deterrent level of a liability claim can be defined as the capacity to dissuade a potential tort-feasor from behaving carelessly and depends on the one hand on the size of a sanction and on the other hand on the probability of being sanctioned (Baker et al. 2004). In combination, these two variables constitute the expected sanction, and the expected sanction is what influences behavior. The more the public authority believes that it will be sued and punished in a way that has a significant impact, the greater the deterrent effect. The third condition that needs to be met is the ability to conform conduct to the requirements of the rules as actors make decisions about their activities (Robbenolt and Hans 2016).

Various circumstances, often arising from the context in which actors operate, can influence these conditions resulting in either over- or under-deterrence. In a situation of under-deterrence, the deterrent capacity of liability is too low to motivate parties to take optimal care while exercising their activities. In other words, potential tort-feasors do not internalize the full costs of their behavior. The opposite occurs in a situation of over-deterrence. In this case, the fear of being held liable is so severe that a potential tort-feasor engages into so-called defensive conduct. Defensive conduct refers to any act or omission that is performed solely to avoid liability or to provide a good legal defense against a liability claim because of over-deterrence (Hauser et al. 1991).

To determine whether liability can adequately deter undesirable activities, encourage defensive conduct, or has no impact at all, it is necessary to take a closer look at the specific context in which financial supervisory authorities operate.

Willingness to Follow the Rules: The Probability of Being Caught, Sued, and Held Liable

The higher the probability of being caught and sued, the more likely it is that liability will deter potential tort-feasors. So, what is the probability that a financial supervisory authority is caught, sued, and held liable by the court for careless behavior resulting in damage? To answer this question, we need to make a distinction between two categories of financial supervisory liability (Tison 2003).

First, financial supervisory authorities can be held liable by third parties. In most cases, this will happen after the bankruptcy of a financial institution. Creditors of the financial institution will then try to receive compensation for their losses by suing the financial authority based on shortcomings in performing their tasks. Without a bankruptcy of a financial institution, the financial supervisory authority does not face a big risk of being held liable as the victims will first try to get compensation from the primary wrongdoer, the financial institution itself. The probability of third-party financial supervisory liability depends thus mainly on the probability of a financial institution going bankrupt.

Under normal circumstances, the chances of a financial institution going bankrupt are relatively small. Even when a financial supervisory authority acts negligent while performing its supervisory tasks, it is not likely that that causes financial institutions to go immediately bankrupt. It merely increases the probability of defaults in the future. The opposite is also true. Despite thorough financial supervision, there is still a possibility that a financial institution goes bankrupt. The bankruptcy of a financial institution depends namely on a wide variety of factors and circumstances of which financial supervision is merely one. The chances of an actual bankruptcy are further reduced by the fact that governments will most likely intervene when a financial institution is "too big to fail." In the past, we have seen that governments will nationalize these troubled financial institutions.

But even if a financial institution goes bankrupt, there is a mechanism that limits the liability risk for a financial supervisory authority. Most countries, namely, have a deposit guarantee system in place that offers compensation to depositors in case a financial institution goes bankrupt (Dijkstra 2009). Only if their losses are (much) greater than the compensation from the deposit guarantee fund will they have an incentive for holding the financial supervisory authorities liable. This group of persons is likely to be limited as

a deposit guarantee fund, with a minimum coverage level of EUR 100,000, normally covers 95% of all deposits (Joint Research Centre 2010). We can thus conclude that the risk of third-party financial supervisory liability is limited, even if financial supervisory authorities are performing their tasks in a negligent manner.

The second category of financial supervisory liability seems to be more straightforward as it concerns a direct relationship between the financial supervisory authority and the victim (e.g., the financial institution subject to supervision). In this case, it is more likely that negligent behavior of a financial supervisory authority results in damage that is immediately detected by the victim. Liability in this category will mainly arise from wrongful actions, in most cases acting too strictly, aimed at the financial institutions. Examples include the wrongful rejection of a permit to operate (market entrance) and a wrongfully published sanction resulting in (reputation) damage for the financial institution. Compared to third-party liability due to too lenient supervision as discussed above, the probability of being held liable by financial institutions due to too stringent supervision is larger.

For both liability categories, it is important to mention that they are affected by the so-called availability heuristic. This behavioral economics term refers to a mental shortcut that relies on immediate examples that come to a person's mind when thinking about a specific topic. Judgments about the probability of being held liable are thus often affected by whether a recent event comes readily to mind. The key issue is thus not the objective risk of being sued and held liable but the perception of the severity of the risk that this will happen (Baker et al. 2004). That might also explain why the Dutch financial supervisory authorities changed their mindset about the threat of liability leading to defensive conduct after a few lawsuits arising from the financial crisis (Dijkstra 2017).

Willingness to Follow the Rules: The Size of the Sanction

As a general starting point, damages should fully compensate the victim for his losses because only then will the injurer internalize the negative externalities that he has caused (Visscher 2009). So, to provide financial supervisory authorities with the correct incentives to behave carefully, damages should be based on the social losses caused by their negligent behavior. A number of factors influence the size of the sanction that financial supervisory authorities face when being held liable.

First, there is the earlier mentioned deposit guarantee system. This system compensates a significant part of the damage of depositors in case a financial institution goes bankrupt. Consequently, financial supervisory authorities will not face the full costs of their careless behavior. The existence of a deposit guarantee scheme is thus likely to decrease the deterrent level of third-party financial supervisory liability (Dijkstra 2009).

Secondly, financial supervisory authorities might have specific clauses in place that shield them from paying the full amount of damages themselves. These clauses often relate to all categories of financial supervisory liability. The Dutch financial supervisory authorities have, for instance, a safeguard clause in place with the Dutch Ministry of Finance. This safeguard clause limits their financial risk from liability claims to maximum 10% of their budget. In case a liability claim exceeds this maximum, the Dutch Ministry of Finance will cover the remainder of the damage compensation. This means that the Dutch financial authorities will then not internalize the full amount of the damage they have caused.

Although not all financial supervisory authorities have formal safeguard clauses in place, it is likely that implicit safeguard clauses exist. It is not realistic to assume that governments would endanger financial supervision by allowing liability claims to fully consume the budgets of their financial supervisory authorities. Furthermore, the fact that financial supervisory authorities are more and more funded by fees imposed on the financial institutions subject to supervision makes it less likely that the authorities will face the full monetary consequences of their careless behavior as they can pass their costs on to the financial industry.

In addition, financial supervisory authorities may also have the possibility to insure the

financial risks from liability claims, as is the case with the Dutch financial supervisory authorities. In general, law and economics theory predicts that liability insurance will reduce the incentives for potential tort-feasors (Cooter and Ulen 2016).

Awareness and Understanding of the Rules

Potential tort-feasors must be aware and understand the legal rules in order for liability to deter negligent conduct. The traditional economic model assumes that there is no uncertainty regarding the level of care needed to be compliant with the legal rules (Cooter and Ulen 2016). In the real world, however, legal standards are often uncertain. To determine the optimal level of due care, courts need complete and accurate information on the costs of care and the expected costs of accidents for each level of care. However, data necessary to set the optimal level of care will often be unavailable. In addition, courts will not always be able to properly observe the actual level of care exercised by the tort-feasor due to measurement errors, insufficient evidence, and misrepresentation about the actual level of care. Thus, only during a trial, when parties present the facts of the case, is the due level of care established in a more precise way. As a result, tort-feasors exercising a certain level of care might not know ex ante whether or not they will be found negligent.

This also applies to financial supervisory authorities. From case law and literature, it becomes clear that the standard for financial supervisory authorities under a negligent liability rule is "reasonable care." This standard is surrounded by uncertainty, because financial supervisors have discretionary powers to fulfill their supervisory duties and face the difficult task of considering both the interests of the supervised institutions and the individual members of society. Furthermore, there are relatively few tort judgements in the area of financial supervisory liability which makes it difficult to know the exact content of the standard of care. What does this mean for the deterrent effect of tort law?

Uncertainty regarding the level of due care changes the deterrent impact of legal rules by creating two opposing effects (Craswell and Calfee 1986). The first effect is an incentive to over-comply. Tort-feasors will, in this situation, take more care than is required by the legal standard of care to increase the chance that they will not be held liable. However, uncertainty also creates a chance that a tort-feasor will not be held liable, thus reducing the incentives to comply with the legal standard. The question thus created is which effect will prevail. Standard law and economic theory predict that over-deterrence (and thus defensive conduct) will occur (Craswell and Calfee 1986). Scholars have argued that this is even worse when a public authority is involved due to the fact that, unlike private tort-feasors, a public authority typically balances two external costs, as it does not bear the costs of over-precaution. Public authorities are therefore much more quickly inclined toward taking excessive care (e.g., De Geest 2011; De Mot and Faure 2012).

The existence of uncertainty regarding the standard of due care for financial supervisory authorities might explain why politicians and many others fear over-deterrence and thus defensive conduct on the side of financial supervisory authorities. By limiting the liability of financial supervisory authorities to cases of gross negligence and/or bad faith, they argue that the standard of care becomes clearer and hence would result in limiting over-deterrence and thus defensive conduct.

Ability to Conform Conduct

The third and last condition that needs to be met in order for liability to effectively deter is the ability of parties to conform their conduct to the requirements of the rules as they make decisions about activities in which to engage, the extent and location of those activities, and any precautions to undertake (Robbenolt and Hans 2016).

Law and economic scholars view organizations and thus also financial supervisory authorities as single economic and organic entities. However, it is important to note that public authorities themselves do not commit negligent acts; their employees do. Financial supervisory liability involves the imposition of liability on the organization itself due to the harmful behavior of its employees. Making the financial supervisory authority liable may however fail to provide

its employees with sufficient incentives to act carefully, as the sanctions imposed on the organization might not reach the responsible employees (Dijkstra 2009; Dari Mattiacci et al. 2010). The question of whether incentives are transferred from the financial supervisory authority to its employees is a manifestation of the well-known agency problem between organizations and their employees. The challenge for financial supervisory authorities is, therefore, one of overcoming principal–agent problems.

While most of the financial supervisory authorities have mechanisms in place to mitigate these problems (e.g., reward policies, recruitment policies, internal and external audits), it is realistic to assume that principal–agent problems dilute, to some extent, the deterrent effect of financial supervisory liability.

Over-deterrence, Under-deterrence, or No Effect at All?

The various factors from the previous paragraphs are more likely to result in under-deterrence than in over-deterrence. While a vague standard of care is likely to result in over-deterrence, several other factors contribute to under-deterrence. First, the threat of third-party liability seems to be limited as this liability category follows, in most cases, the bankruptcy of a financial institution which does not occur often. Second, in case a financial institution goes bankrupt, the existence of a deposit guarantee scheme limits the potential damage of third parties and thus also the potential number of claimants. Third, the existence of explicit or implicit safeguard clauses and the possibility to insure the financial risk will further limit the financial consequences for the negligent behaving financial supervisory authority mitigating the deterrent effect of liability. Furthermore, it is questionable whether the financial supervisory authority can effectively transfer the incentives from liability to their employees. Under-deterrence is therefore most likely to occur, thereby questioning the risk of defensive conduct.

It is however difficult, if not impossible, to accurately estimate the deterrent impact of financial supervisory liability from a theoretical perspective. Whether the imposition of liability promotes more effective financial supervision, encourages defensive practices, or has no discernible effect is at heart an empirical question. The next paragraph presents therefore the outcome of empirical research on this topic.

Empirical Evidence

There is hardly any empirical research on the deterrent impact of financial supervisory liability available. The few studies that exist do show however that defensive conduct is not likely to occur, making under-deterrence more realistic.

Van Dam (2006) asked several Dutch national supervisory authorities, including the financial supervisory authorities, whether their behavior is influenced by the fear of liability. At that time, the supervisory authorities claimed that they did not change their policy out of fear for liability claims. Based on their statements, Van Dam concluded there was no indication for defensive conduct. A couple of years later, the financial supervisory authorities, most likely due to the influence of increased liability claims because of the financial crisis, seemed to have changed their minds and argued for a limitation of their liability. The attitude of the Dutch financial supervisory authorities toward liability claims changed since the financial crisis had put them more into the spotlights. This could indicate that perceptions regarding liability are likely to change when confronted with more liability claims that generate (negative) publicity.

In another study, Trebus and Van Dijck (2014) carried out interviews with four senior employees of the Dutch Financial Markets Authority to examine the impact of liability on their behavior. None of the respondents mentioned any form of defensive behavior in response to liability claims nor did they feel threatened by liability in the exercise of their daily supervisory activities. Based on their research, Trebus and Van Dijck concluded that financial supervisory liability has almost no impact on the behavior of financial supervisors.

A more comprehensive empirical research was carried out in 2015. Dijkstra (2017) conducted a

survey among 500 financial supervisors active in financial supervisory authorities in the member states of the European Union. The majority of the respondents classified the impact of financial supervisory liability as either neutral or positive. The findings of this study therefore imply a modest degree of deterrence. Furthermore, the relative neutral or positive attitude suggests that financial supervisors do not consider financial supervisory liability a burden for executing effective financial supervision. In addition, the survey did not find differences between those respondents who perceive the liability of their organization as limited and those who do not. This suggests that limiting financial supervisory liability does not affect perceptions of the impact of financial supervisory liability or possibly even the behavior of financial supervisors. Therefore, the study calls into question the widely accepted argument of defensive conduct as a reason for limiting the liability of financial supervisory authorities.

This limited empirical research can, however, not be seen as overwhelming empirical evidence regarding the impact of financial supervisory liability. One could further argue that the value of these types of studies (surveys) is limited because they measure perceptions and not actual behavior. They do however raise serious doubts on whether financial supervisory liability will result in over-deterrence and thus defensive conduct.

Conclusion

The theoretical literature is divided regarding the deterrent impact of financial supervisory liability. The picture that emerges from this entry, while lacking detail in many spots, should at least inspire skepticism about the risk of over-deterrence. Our theoretical analysis shows that there are many factors influencing the deterrent level of financial supervisory liability. Most of them tend to lead to under-deterrence, making it hard to believe that financial supervisory liability will result in over-deterrence. This outcome is supported by limited empirical evidence showing only a modest degree of deterrence.

References

Athanassiou P (2011) Financial sector supervisors accountability: a European perspective, European central bank legal working paper series, No. 12

Baker T, Harel A, Kugler T (2004) The virtue of uncertainty in law, an experimental approach. Iowa Law Rev 89:443–487

Cooter R, Ulen T (2016) Law & economics, 6th edn. Addison-Wesley, California

Craswell R, Calfee JE (1986) Deterrence and uncertain legal standards. J Law Econ Org 2(2):279–303

Dari Mattiacci G, Garoupa NM, Gomez-Pomar F (2010) State liability. Euro Rev Private Law 18(4):773–811

de Mot J, Faure M (2012) Report for the European group on tort law: state liability: an economic analysis. Ghent University Academic Bibliography, Ghent

Dijkstra RJ (2009) Liability of financial regulators: defensive conduct or careful supervision? J Bank Regul 11(2):269–284

Dijkstra RJ (2010) Accountability of financial supervisory agencies: an incentive approach. J Bank Regul 3(3):346–377

Dijkstra RJ (2012) Liability of financial supervisory agencies in the European Union. J Euro Tort Law 3(3):346–377

Dijkstra RJ (2017) Is limiting liability a way to prevent defensive conduct? Eur J Law Econ 43(1):59–81

Faure MG (2008) Calabresi and Behavioural tort law and economics. Erasmus Law Rev 1(4):75–102

Geest G de (2011) Who should be immune from tort liability? Washington University in St. Louis Legal Studies Research Paper No. 11-02-04

Gold RM (2016) Compensation's role in deterrence. Notre Dame Law Rev 91(5):1997–2048

Hauser MJ, Commons ML, Bursztajn HJ, Gutheil TH (1991) Fear of malpractice liability and its role in clinical decision making. In: Gutheil TG, Bursztajn HJ, Brodsky A, Alexander V. (eds) 1991Decision making in psychiatry and the law. Williams & Wilkins Co, Baltimore, pp 209–226

Joint Research Center (2010) Impact assessment to the commission proposal for the recast of the Deposit Guarantee Schemes Directive

Kramer L, Sykes AO (1987) Municipal liability under § 1983: a legal and economic analysis. Supreme Court Rev:249–301

Lastra RM (2006) Legal foundations of international monetary stability. Oxford University Press, Oxford

Levinson D (2000) Making government pay: markets, politics and the allocation of constitutional costs. Univ Chicago Law Rev 67:345–420

Niskanen WA Jr (1971) Bureaucracy and representative government. Aldline, Atherton, Chicago

Robbenolt JK, Hans VP (2016) The psychology of tort law. New York University Press

Rosenthal L (2006) A theory of governmental damages liability: torts, constitutional torts, and takings. J Const Law 9:1–73

Spitzer ML (1977) An economic analysis of sovereign immunity in tort. South Calif Law Rev 50:515–548
Tison M (2003) Challenging the prudential supervisor: liability versus (regulatory) immunity, Financial Law Institute Working Paper 2003-04
Trebus J, van Dijck G (2014) Effecten van aansprakelijkheid op het handelen van de AFM empirisch onderzocht (Empirical research of the impact of liability on the Dutch financial markets authority). Aansprakelijkheid, verzekering en schade 14(4): 99–107
van Dam CC (2006) Liability of regulators. An analysis of the liability risks for regulators for inadequate supervision and enforcement, as well as some recommendations for future policy. British Institute of Comparative and International Law, London
Visscher LT (2009) Tort Damages. In: Faure MG (ed) Tort law and economics, encyclopedia of law and economics. Edward Elgar, Cheltenham, pp 153–200

External Incentives

► Intrinsic and Extrinsic Motivation

Externalities

Andrew Torre
School of Accounting, Economics and Finance, Deakin University, Melbourne, VIC, Australia

Abstract

Externalities and court output are closely related since the primary economic function of the courts is to price and thereby internalize external effects in missing markets. This is achieved using the complementary technologies of torts and criminal law. An externalized cost can be broken up into two components, the cost imposed on the unconsenting victim and the cost imposed on wider society, such as insecurity and fear of further random episodes of the same event. Damages awarded to the victim in torts, and criminal punishment inflicted upon the offender, value and internalize the first and second components, respectively. Unless damages and punishment accompany each other, the externality will not be fully internalized. In addition, punishment produces absolute general deterrence, which is a positive externality, and in this sense negative and positive externalities are mirror images of each other, or perfect complements, in a legal economics framework. This framework also establishes the theoretical underpinnings and basis for valuing joint court output in the civil and criminal jurisdictions.

Synonyms

Externalized benefits; Externalized costs

Definitions

Externalities are classified as either negative or positive.

A negative externality arises when a decision maker does not bear all of the costs of his or her decision. As a consequence, some costs are externalized onto others without compensating them. An example would be burglary. The private cost to the burglar is the dollar value of the alternative use of his or her time and the expected penalty if caught and convicted. However, the social cost is much higher. First, there is the cost imposed on the victim in the form of property damage and/or loss, and in addition, intangible costs such as fear of further incidents and distrust of others, leading to excessive precautions being taken to protect the property. Second, people other than the victim also incur the fear and self-protection costs. The social cost of burglary therefore exceeds the private cost.

A positive externality arises when a decision maker does not capture all of the benefits of his or her decision. If the burglar is caught and convicted, then a judge will impose punishment in the form of a monetary, noncustodial or custodial penalty. The sanction produces a positive externality in the form of absolute general deterrence, because punishing the burglar deters others from similar behavior in the future. There is a private benefit to the victim, since the penalty may prevent burglary of his or her property in the

future and a much wider benefit to everyone else for the same reason. Therefore, the social exceeds the private benefit.

Introduction

A useful starting point for this essay is to conceptualize the legal system as a production process. In economic theory, a production function summarizes the relationship between the inputs, labor and capital, and the corresponding outputs for a given technology. Similarly, the legal system utilizes a given technology to produce several intermediate outputs and two distinct final outputs from a bundle of inputs. Lawyers distinguish between civil and criminal cases. A case and its ingredients constitute the raw materials of the legal system. For a civil matter these are the cause of action(s), the nature and quality of the evidence and the services of solicitors and barristers, and in a criminal case the offense(s), police, prosecutor, solicitors, and barristers' services. The two final joint civil and criminal court outputs that can be valued in dollar terms are damages and absolute general deterrence, respectively. The latter is jointly produced by the police and courts. Intermediate civil outputs include the court's finding for the plaintiff or defendant, orders of specific performance, injunctions, and in the case of family law matters, custody and property division decisions, for example. Intermediate criminal outputs are bail decisions, committals, trials, sentencing hearings, and the verdict. Absolute general deterrence means that as a consequence of the police clearing up offenses and courts inflicting punishment in the form of fines and jail time on convicted defendants, the social cost of crime will be lower in the future. This is because some prospective offenders at the margin of legal and illegal activity will desist from the latter, since the risk of expected punishment does not make it worthwhile. For this reason, punishment that deters is sometimes called utilitarian punishment (Bagaric 2000). Society is unequivocally better off and therefore net happiness increases. It would only be the case that punishment never deters some people other than the offender, if penalties for serious offenses were abolished and their incidence did not increase. This seems highly improbable.

The technology of the legal system comprises three interrelated parts. First in common law systems, there is a body of complex past decisions called precedents or more generally the common law. Its birth approximately occurred with the establishment of the royal courts or curiae regis by Henry 11, who reigned from 1154 to 1189 (Mendelson 2010, p. 12). Second, the stock of judicial precedents is supplemented by a voluminous compendium of statutory-based law. The equivalent in civil law systems is a set of codes, modeled on the classical Roman law. However, unlike in common law jurisdictions, codes in civil law ones are the most authoritative source of law with judicial decisions being subservient to them. Third, there are the judges who create, interpret, and apply the law.

A production process is said to be efficient if the cost of using the inputs to produce the outputs is minimized; the same reasoning applies to the legal system; damages and deterrence can be supplied efficiently or inefficiently, for a given technology. However, this use of the term efficiency is generally different from the way it is used in the law and economics literature. In the latter, it tends to be employed by some writers when analyzing bodies of substantive law, in particular torts, criminal, contract, and property law (Landes and Posner 1981, 1987; Cooter 1985; Posner 1985, 2010). According to this paradigm, torts legal rules, for example, are efficient if they prospectively minimize the sum of expected damage plus precaution or avoidance costs (Cooter and Ulen 2004, Kaplow and Shavell 1999, Posner 2010). This hypothesis has stimulated the production of a large critical literature. For example, applying it to legal negligence would necessitate the court engaging in a fairly complex calculation exercise, in its inquiry as to whether the injurer had taken cost-effective precautions against the legally caused harm. Hayek's fundamental contribution to the socialist calculation debate in the 1920s and 1930s, in which he argued vigorously against central planning as a substitute for the market mechanism in allocating scarce resources between

alternative uses, was to show that localized, widely dispersed information precluded rational economic calculation on the part of economic planners (Hayek 1945). Similarly, a judge asked to undertake an efficiency cost-benefit calculation would find him- or herself in the same position as a central planner. This led Leoni and Hayek to an alternative characterization of the legal system, particularly common law dynamics, as a spontaneous order derived from the adjudication of individual claims and not the result of human design that aims at any particular end such as wealth maximization or efficiency (Leoni 1972; Aranson 1988; Hayek 1973, Volume 1; Cheren 2012). It seems that this approach fits more comfortably with rights based non-efficiency theories such as corrective justice, which Burrows (1999) defines as "protection of arbitrary alterations to the initial distribution of property rights by restoration of plaintiffs as far as possible to their ex-ante position, thereby preventing defendants from benefiting from the actual harm they have caused."

The Economic Nature of Externalities and Their Relationship to Law

The subject matter of this essay is externalities, which can be broadly classified into negative and positive (Gans et al. 2012). Textbooks commonly use pollution as an example of the former, and research and development expenditure is common for the latter. A firm that produces steel incurs private costs of $100,000 (raw materials and labor costs). Residents living across the road suffer from asthma as a consequence of pollution from the factory and incur medical costs of $10,000. While the total social cost of producing steel is $110,000, market forces will only reflect the private cost of $100,000, since the manufacturer externalizes $10,000 on to the residents. Consequently, market forces overvalue steel and too much is produced and consumed. As the fundamental cause of the problem is an absence of enforceable property rights in the airspace through which the pollution is transmitted, property law provides the basis for a market or private solution. If either a court or the legislature vests property rights in the airspace to the firm, then the affected residents may be able to bribe the firm to cut back its emissions of pollution in return for a payment. Alternatively if the residents own the airspace, then the firm may be able to successfully negotiate to pay the residents to accept a mutually agreed upon level of pollution. Irrespective of the initial assignment of property rights, transaction costs may impede a bargain from being successfully consummated (Cooter and Ulen 2004; Posner 2010). Sources of high transaction costs include emotional or psychological conflict between the parties, information about the damage cost not being fully available to both parties, disagreements about the cost estimates making it difficult to negotiate a price, or too many parties to deal with (Cooter and Ulen 2004; Posner 2010). The more people that are involved, the less likely it is that an agreement acceptable to everyone will be reached. If bargaining fails, then the standard solution advocated is to force the firm to take into account the $10,000 externalized cost when it is computing its profit, by imposing a tax or a fine equal to this amount (Gans et al. 2012). This is only efficient if it would be more expensive for the affected residents to take corrective action themselves, for example, by relocating.

In contrast to pollution, research and development expenditure produces a positive externality; consequently market forces will undersupply it. Unimpeded market forces will lead private firms to underinvest in knowledge creation because they are unable to fully appropriate the benefits of their investment. For example, in the case of a new drug that is very expensive to develop, it would be relatively easy for a competitor to discover its molecular formula through reverse engineering. As a consequence drugs will be undersupplied and overvalued (priced too high) by market forces. Traditional solutions are government subsidies, and patenting of the intellectual property embodied in the drug, to rectify the market failure.

These two examples suggest some sort of relationship between legal technology, in particular physical and intellectual property law, and externalities; however, the relationship is at best

cursory. Yet using a legal economics framework, it is possible to see the centrality of torts and criminal law to the problem at hand, and their relationship to the two externality types, in a different light. Torts and criminal law are very much concerned with defining and internalizing negative externalities, and in addition, perhaps counterintuitively, criminal law generates a positive externality. An externalized cost can be broken up into two components, the cost imposed on the unconsenting victim, for example, violation of bodily integrity following a sexual assault, and the cost imposed on wider society, for example, insecurity and fear of further random episodes of the same event. A finding in favor of the victim, followed by an award of damages in torts, values and internalizes the first, while a guilty verdict, followed by punishment, values and internalizes the second component. Full internalization occurs because the injurer is forced to confront both components of the externalized cost. As well, punishment produces deterrence, which is a positive externality or externalized benefit. Since an award of damages without a criminal penalty will only achieve partial internalization, economically, torts and criminal law are complementary technologies, not substitutes.

Historically, this was not always the case, because there were alternative ways for a victim to pursue justice for the same wrongful act, a choice between "compensation or vengeance" (Seipp 1996). If compensation was chosen, it would be paid to the victim and the king, so that the externalized cost would be fully internalized. However, most victims tended to choose vengeance since not many wrongdoers had sufficient wealth to pay the necessary compensation and during the Middle Ages (fifth to the fifteenth century), revenge was considered a "higher right" (Frankel 1996). Professor Mendelson classifies modern torts into three categories: statutory, trespass, and action on the case (Mendelson 2010). Trespass comprises battery, assault, false imprisonment, variants of statutory trespass, trespass to land, trespass to goods, and cattle trespass, while case comprises, inter alia, deceit, detinue (unlawful deprivation of possession), conversion, nuisance, libel, slander, defamation, negligence, passing off, and a series of miscellaneous intentional torts of action on the case for personal injury (idem. p. 8). These causes of action protect a multiplicity of entitlements, which include physical integrity or immunity from direct and indirect injury (battery and negligence); the right to use land, light, air, running water, the sea, and its shore (trespass to land, private and public nuisance, and negligence); and the right to corporeal and intellectual property (conversion, detinue, trespass to goods, passing off, misrepresentation, and injurious falsehood) (idem. p. 6). All of these causes of action have equivalent counterparts in the criminal law.

Two Illustrations of the Close Nexus Between Externalities and Torts and Criminal Law

Two detailed examples are now used to illustrate the themes exposited in this essay. The first is the case of a property owner who proposes buying an airspace lot above the roof of an apartment block, which is adjacent to his top floor apartment. The entire apartment block is owned by one person and the objective of the property owner's purchase is to preserve his unimpeded city views. If this market transaction fails and the adjacent owner in any way obstructs the property owner's view, she will externalize an uncompensated cost onto him. As in any market transaction, the property owner will have a maximum willingness to pay for unimpeded views and the apartment owner will have a minimum or reservation price that she will require before relinquishing the right to her air space lot. For a deal to be consummated, the negotiated price will need to lie between these two values. The two most likely stumbling blocks to a successful outcome would be agreeing on a mutually acceptable price and/or animosity between the parties. These potential problems are called transaction costs and can prevent the completion of a successful bargain that would internalize the externality (Cooter and Ulen 2004; Gans et al. 2012; Posner 2010).

If bargaining fails and the apartment owner subsequently erects a structure on her roof without

in anyway invading the property owner's airspace, then even though she is externalizing a cost on to him, no common law legal right would have been violated. This is because, as reaffirmed by the English House of Lords in 1997 in the case of Hunter v Canary Wharf Ltd., the law of torts does not recognize a legal right to a view from one's property (Mendelson 2010, p. 668). If however the property owner ignored his neighbor's refusal to sell her air space lot and attempted to coerce her into changing her mind, by, for example, entering the balcony of the neighbor's top floor apartment without her permission, then this would infringe a legally recognized right. This externalized cost would be actionable at common law for trespass to the neighbor's land (Mendelson 2010, p. 138). While an injunction ordering the tortfeasor to desist from the trespass would enforce the neighbor's property right, it would not in the economic sense force the property owner to confront the full costs of his decision to bypass the market. The first step in achieving this, i.e., fully internalizing the externality, would be a finding for the victim and an award of compensatory damages, which measures the court's assessment of the externalized cost, imposed on the neighbor. The second step is a finding of guilt and the infliction of punishment. For example, section 9(1)(e) of the Victorian Summary Offences Act 1966 provides for a fine of $3,500 or up to 6 months imprisonment for criminal trespass. In the absence of punishment, the rest of society would not be compensated for the costs imposed on it. Those social costs include apprehension, distrust of others, and excessive expenditure on self-protection. The deterrence that follows punishment, as well as being an externalized benefit or positive externality, is a public good, because everyone in society is able to consume it simultaneously and no one can be excluded on the basis of willingness and ability to pay. Punitive or exemplary damages would be an imperfect substitute for punishment, even though it is unlikely that they could be insured against, because as already noted, the ability of the injurer to pay can substantially impact their deterrence value. This is particularly the case in countries such as Australia where a defendant cannot be punished twice by an award of punitive damages and criminal conviction (Mendelson 2010, p. 48).

Legal negligence provides the background for the second example. Mendelson writes: "the advent of railways in the early 1830s, which brought in its wake an unprecedented toll of accidental injuries and death, provided the impetus for the development of negligence as a separate tort" (Mendelson 2010, p. 277). The entitlement not to be injured or killed by negligent driving is one of the rights recognized and protected by this tort and the criminal law. Protecting road and street users' rights not to be interfered with by negligent driving using the market mechanism or Coase bargains would entail very high transaction costs and thus market failure (Coase 1937, 1960). Drivers would have to locate and then contact every potential victim of their negligent driving and then negotiate a price at which each victim would be prepared to relinquish the right not to be negligently harmed. Alternatively, every potential victim would have to locate and contact every potential negligent driver, and then negotiate the price to buy the right not to be harmed. For both injurer and victim, it would not be economical to use cars, streets, or the roads because the search and bargaining costs, (transaction costs) would be too high (Posner 2010). In this case however, unlike the first where the property owner is compelled to bargain with his neighbor if he does not want his view to be obstructed, transaction costs are too high to justify a market solution.

Legal negligence requires the victim to establish that the injurer breached a legally recognized duty of care, which factually causes reasonably foreseeable damage (Mendelson 2010, p. 281). Similar elements constitute the criminal counterpart of the tort of negligent driving. For example, S.318(2)(b) of the Victorian Crimes Act 1958 provides that:

> Any person who by the culpable driving of a motor vehicle causes the death of another person shall be guilty of an indictable offence and shall be liable to level 3 imprisonment (20 years maximum) or a level 3 fine or both. For the purposes of subsection (1) a person drives a motor vehicle culpably if he drives the motor vehicle negligently, that is to say, if he fails unjustifiably and to a gross degree to

observe the standard of care which a reasonable man would have observed in all the circumstances of the case.

In both of these examples, the cause of the externalized cost problem is not an absence of property rights, since these are well defined in both cases. The respective sources of the problem are market bypass where a market transaction would have been relatively cheap and failure to observe a legal standard of care in the face of prohibitively high transaction costs that make market bypass impossible (Posner 2010).

Judicial Valuation of Externalities

The notion of "price lists" for externalized harms can be found very early on in the history of Anglo-Saxon and Germanic law. An example is "the code of Æthelberht, the King of Wessex, which contains elaborate tariffs of fines for breach of the peace" (Mendelson 2010, p. 10). More generally, "wergeld" tables served two functions. First, they provided for a fixed scale of compensation, which was determined by the victim's social and legal status. For example, the "wergeld" of princes and free lords was 360 shillings, the judicial class 30 shillings, rent-paying tenants and other freemen 15 shillings, and a day laborer's "wergeld" was paid in wheat (Mendelson 2010, p. 34). Second, the tables put a monetary value on injuries to different parts of the body, which was proportional to the victim's "wergeld"; for example, the price of permanent injury to the mouth, nose, eyes, tongue, ears, male sexual organs, hands, and feet was set equal to one half of the victim's "wergeld" (Mendelson 2010, p. 35).

In contemporary times judicial valuations of externalized costs can be inferred from the average quantum of compensation they award in torts to victims and the average quantum of punishment they impose on offenders. The averages aggregate and summarize the dispersed or localized information about the distribution of diverse cases for a particular cause of action and offense. They are socially optimal because the judges' comparative advantage is in producing justice according to law. This requires access to all of the relevant information about the case, and knowledge of the relevant substantive law, which they only possess. The last statement needs to be somewhat qualified when considering the 'optimality' of court awarded tortious damages. In Australia, following reform legislation early this century, a distinction is now made between intentional and unintentional wrongs; the former are still governed by the common law of damages, while the latter are governed by statute (Mendelson 2010). Statutory based damages are subject to statutory thresholds and capping (Mendelson 2010). Consequently, judicially determined awards in the case of negligence for example, are constrained optima. However common law principles also considerably constrain the courts' assessment of the victim's disability. In estimating 'the quantum of damages for past and future pain and suffering, loss of enjoyment of life and loss of earning capacity' courts apply the 'once for all' rule established in 1699 (Mendelson 2010). Compensation is a lump sum, which cannot be subsequently changed, if the plaintiff's injury worsens or develops into something more serious down the track (Mendelson, 2010). Similarly, when victims sustain very serious and permanent injuries, for example quadriplegia, courts face insuperable obstacles in forecasting the victim's future loss of earning capacity and nursing requirements. As noted by Mendelson (2010), this situation makes it very likely that the gravity of the victim's injury and injurer advantage are positively correlated. Average damages denoted by AD measure the price of the externalized harm borne by the victim. The supply of absolute general deterrence plus the judicial valuation of the externalized cost to society implied by the optimal punishment is found by minimizing Eqs. 1 and 2 with respect to X* and F*, the optimal jail sentence and fine, respectively:

$$\begin{aligned} \mathrm{MIN.E[C]} &= \mathrm{E[D[O[X^*]]]} + \mathrm{P}\,X^*\,\mathrm{b}\,\mathrm{O[X^*]} \\ &= \mathrm{D[O[X^*]]} + \mathrm{EP}\,X^*\,\mathrm{b}\,\mathrm{O[X^*]} \end{aligned} \tag{1}$$

$$\begin{aligned} \mathrm{MIN.E[C]} &= \mathrm{E[D[O[F^*]]]} + \mathrm{P}\,F^*\,\mathrm{O[F^*]} \\ &= \mathrm{D[O[F^*]]} + \mathrm{EP}\,F^*\,\mathrm{O[F^*]} \end{aligned} \tag{2}$$

where E[C] is expected cost, D is harm, O is offenses, EP is the expected probability of the offender being jailed or fined, b is the cost of incarcerating an offender each time period, and X* and F* are the average jail sentence and fine for the offense. The asterisk indicates that they are exogenous and optimal in the sense explained earlier. The solutions are given by Eqs. 3 and 4, respectively:

$$\frac{F^{*}EP}{e} + F^{*}EP \tag{3}$$

$$\frac{X^{*}b^{*}EP}{e} + X^{*}b^{*}EP \tag{4}$$

The first term on the RHS of Eqs. 3 and 4 is the value of the positive externality corresponding to the optimal fine F* and term of imprisonment X*, given EP the probability of the sanction being imposed and e the absolute general deterrence elasticity, whose value will lie between 0 and −1. A value of −0.5, for example, indicates that a 1% (10%) increase in the average fine or jail sentence lowers the expected future costs of crime by 0.5% (5%), etc. A value of −1 indicates perfect absolute general deterrence. The second term of Eqs. 3 and 4 is the court's valuation of the externalized social cost of the offense, i.e., the burden placed on wider society:

$$V_N = AD + F^{*}EP \tag{5}$$

$$V_N = AD + X^{*}b^{*}EP \tag{6}$$

Expressions 5 and 6 give the dollar value (V_N) of the court's valuation of the negative externality implicit in the award of tortious damages (AD) and the punishment imposed for the corresponding offense, while Eqs. 7 and 8 give the value of the positive externality (V_P) from deterrence:

$$V_P = \frac{F^{*}EP}{e} \tag{7}$$

$$V_P = \frac{X^{*}b^{*}EP}{e} \tag{8}$$

The two cases discussed in the third section of this essay can be used to illustrate how these expressions would be translated into dollar values. The parameter values used are deterrence elasticity –0.3, the probability of a fine or imprisonment 0.5, and the annual cost of incarceration $80,000.

Case 1

	Damages	Fine (average value)	Imprisonment
Tort of trespass to land	$5,000		
Offense of trespass to land		$5,000	6 months
Judicial value of negative externality			
(i) Damages + fine		**$7,500**	
(ii) Damages + imprisonment			**$15,000**
Judicial value of positive externality			
(i) Fine		**$8,333**	
(ii) Imprisonment			**$33,333**
Total: negative and positive externality		**$15,833**	**$48,333**
Civil output	**$5,000**		
Criminal output		**$10,833**	**$53,333**
(Assumes only 1 case: civil and corresponding criminal matter)			

Case 2

	Damages	Fine (average value)	Imprisonment
Tort of negligence	$20,000		
Offense of culpable driving		$20,000	1 year
Judicial value of negative externality			
(i) Damages + fine		**$30,000**	
(ii) Damages + imprisonment			**$60,000**
Judicial value of positive externality			
(i) Fine		**$33,333**	
(ii) Imprisonment			**$133,333**
Total: negative and positive externality		**$63,333**	**$193,333**
Civil output	**$20,000**		
Criminal output		**$43,333**	**$173,333**
(Assumes only 1 case: civil and corresponding criminal matter)			

Summary

The production of goods and services that are sold in the marketplace increases social welfare because it generates a social surplus, profit plus consumer surplus. The latter is the difference between the average prices paid for an item and buyers' maximum willingness to pay and is a measure of consumer welfare. Profit measures business surplus, which is the difference between revenue and the opportunity cost of all scarce resources used to generate it. Conceptually

valuing the output of products that are priced in the market is straightforward. This is not the case however for nonprofit services that are not priced by the market mechanism.

Final output valuation in these cases requires a well-developed theory of the organization's economic functions. In relation to the courts, the position adopted in this essay is that they price external effects in missing markets, in both the civil and criminal jurisdictions. The valuation of their final output follows logically from this. Average civil and criminal jurisdiction output is equal to expressions 5 + 7, where fines are used as punishment, and expressions 6 + 8, where imprisonment is used. Total civil jurisdiction output is average damages multiplied by the number of cases (the first term in Eqs. 5 and 6), and total criminal jurisdiction output is the second terms of Eqs. 5 and 6 plus Eqs. 7 and 8 multiplied by the number of cases.

As the two examples show, the pricing exercise is related to the sanction used; the negative externality of land trespass is judicially valued at $7,500 (damages + fine) and $15,000 (damages + imprisonment), while the corresponding figures for negligent driving are $30,000 and $60,000, respectively. The value of the positive externality or public good of deterrence is $8,333 (fine), $33,333 (imprisonment) for land trespass, and $33,333 and $133,333 for negligent driving.

References

Aranson P (1988) Bruno Leoni in retrospect. Harv J Law Publ Policy 11:661–711

Bagaric M (2000) Punishment and sentencing: a rational approach. Cavendish Publishing, London

Burrows P (1999) A deferential role for efficiency theory in analysing causation-based tort law. Eur J Law Econ 8:29–49

Cheren R (2012) Tragic parlor pigs and comedic rascally rabbits: why common law nuisance exceptions refute Coases's economic analysis of law. Case West Law Rev 63(2):556–596

Coase R (1937) The nature of the firm. Economica 16(4):386–405

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

Cooter R (1985) Unity in tort, contract and property: the model of precaution. Calif Law Rev 73(1):1–51

Cooter R, Ulen T (2004) Law and economics, 6th edn. Pearson Addison Wesley, Boston

Frankel T (1996) Lessons from the past: revenge yesterday and today. Boston Univ Law Rev 76:89–102

Gans J, King S, Gregory Mankiw N (2012) Principles of microeconomics. Cengage Learning Victoria. Australia.

Hayek F (1945) Use of knowledge in society. Am Econ Rev 35(4):519–530

Hayek F (1973) Law legislation and liberty. Rules and order, Vol 1. The University of Chicago Press, Chicago.

Kaplow L, Shavell S (1999) Economic analysis of law. Working paper 6960. National Bureau of Economic Research, Cambridge, Mass.

Landes W, Posner R (1981) The positive economic theory of tort law. Ga Law Rev 15:851–921

Landes W, Posner R (1987) The economic structure of tort law. Harvard University Press, Cambridge, Mass.

Leoni B (1972) Freedom and the law. Nash Publishing, Los Angeles

Mendelson D (2010) The new law of torts, 2nd edn. Oxford University Press, Australia and New Zealand.

Posner R (1985) An economic theory of the criminal law. Columbia Law Rev 85(6):1193–1231

Posner R (2010) Economic analysis of law, 8th edn. Aspen Publishers, New York

Seipp D (1996) The distinction between crime and tort in the early common law. Boston Univ Law Rev 76:59–88

Externalized Benefits

▶ Externalities

Externalized Costs

▶ Externalities