Imitation

Jen-Te Yao
Fu-Jen Catholic University, New Taipei City, Taiwan

Abstract

Imitation is an idea adopted or derived from an original one and is a necessary tool for disseminating knowledge. The existence of imitation may lower the monetary returns of innovators and remove their creation incentives. This gives the rationale for the intellectual property right (IPR) protection system, whereby the returns of innovators can be secured and the incentives to innovate can be improved. The pertinence of such a viewpoint is, however, in doubt in some of innovative industries, such as software, computers, and semiconductors. Institutions that govern the creation and diffusion of inventions (knowledge) should balance the trade-off between IPR protection and imitating diffusion.

Introduction

Invention creates new knowledge, and a successful invention often leads to widespread imitation of other innovations. Imitation is an idea adopted or derived from an original one and is a necessary tool for disseminating knowledge. Knowledge accumulation and innovation diffusion must be achieved via imitation. While invention is a discovery and proof of the workability from a new idea or method, innovation is the successful application of new inventions into marketable products.

Imitation is related to copying. Copying is an extreme case of imitation, because it is an exact reproduction of the original, such as a fake painting. Sometimes a fake painting can be so real looking that it is hard to tell the fake from the genuine article. One might recall a common proverb that "imitation is the sincerest form of flattery," although artists and museum directors would probably disagree. An excellent invention may signal to imitators that opportunities are "good" and hence implicitly encouraging imitation. Without this signal, potential competitors might have no incentives to imitate.

When discussing imitation, we should understand the difference between imitation, counterfeiting, and piracy. Counterfeiting refers to the illegal copying of trademarks, patent or copyright infringement, where the protected rights to intellectual property are being violated. A counterfeit good is an unauthorized imitation of a branded good. Piracy is making an unauthorized exact copy – not a simple imitation – of an item covered by intellectual property rights (hereafter IPRs).

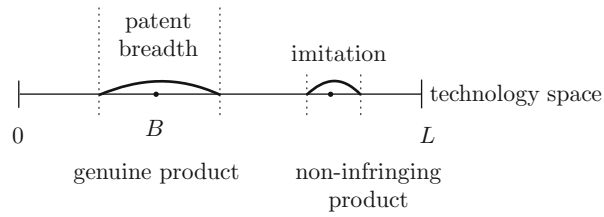
A. Marciano, G. B. Ramello (eds.), *Encyclopedia of Law and Economics*,
https://doi.org/10.1007/978-1-4614-7753-2

Imitation and the Incentives to Invent

Imitation as a factor in economic development has gradually received greater attention. A number of economists have examined the subject. They argued that firms imitating major inventions and diffusing them throughout the economy – not the original inventions themselves – play a critical role in economic growth. This is because complete technological progress consists of three stages: invention, imitation, and diffusion. In the last stage, diffusion is the process of the spreading of inventions through licensing, imitation of patented innovations, or adoption of unpatented innovations. Nobel Prize winner Jean Tirole emphasizes the importance of imitation and diffusion in economic progress. He notes that "Inventing new products is not sufficient for economic progress. The innovations must then be properly exploited and diffused through licensing, imitation, or simple adoption. The difference between imitation and adoption is that imitators must pay for reverse engineering – e.g., figuring out the original firm's technology" (Tirole 1988). Therefore, an imitating work is not free, but rather has a cost.

Inventions and innovations require monetary returns to innovators in a market-based system. However, the existence of imitation may lower the returns and remove the creation incentives of innovators. This gives the rationale for the IPR protection system, whereby the returns of innovators can be secured. In general, stronger IPR protection raises imitation costs and reduces the number of imitators, which can be realized from the mechanism of patent protection. Patent protection has two dimensions: patent length protection and patent breadth protection. Nordhaus (1969) is the first to offer a model dealing with patent life or patent length, i.e., the number of years that the patent is in force and protected from being copied. Nordhaus (1972) extends his model to study the effectiveness of patent breadth protection. Patent breadth indicates the extent to which a given patent covers the field to which it pertains. The breadth of a patent grant measures the degree of protection in the scope of a product's characteristics. For example, if a company invents a new drug to alleviate a heart condition, to what degree can a competitor be allowed to sell a similar drug? If a computer software firm markets a new program, how different should rival products be from the original? One can conclude that if patents are narrow, then they are easy to "invent around"; that is, it is easy to produce a non-infringing substitute for the patented inventions. This constitutes a non-infringing imitation. An extremely narrow patent does not protect a product even against trivial changes, like color or size (Scotchmer 1991).

One can apply a spatial line to express this idea, as illustrated in Fig. 1. Suppose a firm invents a new product and obtains a patent grant. The breadth B of the patent is an exclusion zone in a technology space, whereby the patent holder has exclusive control over the application of the patented technology. The technology space is a closed interval due to the limitation of technology development. Now, a competitor enters the market and needs to consider where to locate its product characteristic legally in the technology space, in order to avoid infringement of the patent. For example, the competitor needs to employ engineers to conduct reverse engineering and to figure out the original firm's technology. Such operations are costly. Imitation entry by a second firm causes the patented firm to compete in the given technology space. However, a broader patent protection leaves a smaller remaining technology space to potential competitors (imitators), making it more difficult for them to enter the



Imitation, Fig. 1 Non-infringing imitation in a technology space

market. One thus concludes that a broader patent leads to less entry and less competition. Consequently, the market price is driven to increase due to less competition, which hurts consumer surplus, but increases the original firm's profits. Imitation entry stops when the cost of imitation can no longer be covered by competition in the market.

Two implications are derived from the above discussion. First, a broader patent increases the original firm's profits due to less competition in the market, thereby improving the incentives of the original firm to engage in innovations. Second, an increase in patent breadth means a rise in imitators' entry cost. Such a view is supported by Gallini (1992). One can further infer that an increase in patent protection will reduce the phenomenon of imitation and thus reduce competition between firms in the market. On the other hand, if a patent is very narrow (and even in some cases where there is no patent protection) and imitation is sufficiently cheap, then firms will prefer to imitate rather than innovate. Under this situation a "waiting game" takes place (Katz and Shapiro 1987) – all firms wait for other firms' inventions to occur, and the result is in fact no innovations. See also Dasgupta (1988) for an early discussion and Choi (1998) who as part of a wider paper on patent litigation, patent strength, and imitation investigates the waiting game style behavior in imitation.

The pertinence of the above viewpoints is, however, much in doubt (Takalo 2001). Ever since the pioneering study by Mansfield (1961), researchers have reported evidence of the inability of patent protection to prevent imitation, with a few exceptions such as in the pharmaceutical industry. Bessen and Maskin (2009) observe an interesting phenomenon that some of the most innovative industries of the last 40 years – software, computers, and semiconductors – have historically had weak patent protection and have experienced rapid imitation of their products. Surveys of managers in semiconductor and computer firms typically report that patents only weakly protect innovation. Patents were rated weak at protecting the returns to innovation, far behind the protection gained from lead time and learning-curve advantages (Levin et al. 1987). Patents in the electronics industries have been estimated to increase imitation costs by only 7% (Mansfield et al. 1981) or 7–15% (Levin et al. 1987).

Imitation and Social Welfare

The strength of IPR protection substantially affects the speed of imitation. The following question arises: Does a strengthening of IPR protection have economic benefits for the society? The empirical literature on innovation shows that "imitation" is a nontrivial exercise that (even in the absence of a patent) may require substantial time and effort (Dosi 1988). Increased protection of IPRs makes imitation more difficult, which generates a trade-off on the improvement of social welfare. On the consumer demand side, stricter IPR protection implies that there will be less price competition between firms, resulting in higher prices and thus a reduction of consumer surplus. However, on the supply side, stricter IPR protection reduces competition between firms, which creates greater returns for the existing innovators and provides incentives for them to innovate. Therefore, stricter IPR protection is bad for consumers' well-being, but is good for innovators' private returns: no one wants to invest in the creation of inventions if imitation cost is very low and dissemination of inventions occurs rapidly.

To summarize, from society's point of view, the welfare effects of imitation depend on the balance of two elements: one is to provide a means for the knowledge producer to capture the benefits of efforts put forth and the other is to maximize the social dissemination of the related inventions. Institutions that govern the creation and diffusion of inventions (knowledge) should balance this trade-off, and this is why it is important to devise social mechanisms to allow inventors to capture a fraction of the benefits generated by the inventions.

Cross-References

▶ Counterfeiting Models: Mathematical/ Economic

References

Bessen J, Maskin E (2009) Sequential innovation, patents, and imitation. RAND J Econ 40(4):611–635
Choi JP (1998) Patent litigation as an information-transmission mechanism. Am Econ Rev 88(5):1249–1263
Dasgupta P (1988) Patents, priority and imitation or, the economics of races and waiting games. Econ J 98:66–80
Dosi G (1988) Sources, procedures, and microeconomic effects of innovation. J Econ Lit 26(3):1120–1171
Gallini N (1992) Patent policy and costly imitation. RAND J Econ 23:52–63
Katz ML, Shapiro C (1987) R&D rivalry with licensing or imitation. Am Econ Rev 77(3):402–420
Levin R, Klevorick A, Nelson R, Winter S (1987) Appropriating the returns from industrial research and development. Brook Pap Econ Act 3:783–820
Mansfield E (1961) Technical change and the rate of imitation. Econometrica 29:741–766
Mansfield E, Schwartz M, Wagner S (1981) Imitation costs and patents: an empirical study. Econ J 91:907–918
Nordhaus W (1969) Invention, growth and welfare. MIT Press, Cambridge, MA
Nordhaus W (1972) The optimal life of the patent: reply. Am Econ Rev 62:428–431
Scotchmer S (1991) Standing on the shoulders of giants: cumulative research and the patent law. J Econ Perspect 5(1):29–41
Tirole J (1988) The theory of industrial organization. MIT Press, Cambridge, MA
Takalo T (2001) On the optimal patent policy. Finn Econ Pap 14(1):33–40

Immigration Law

Thomas Eger and Franziska Weber
Faculty of Law, Institute of Law and Economics, University of Hamburg, Hamburg, Germany

Abstract

In this article we provide some insights into relevant law and economics research on immigration laws and policies. After discussing some typical motives of migrants, we deal with the most important welfare effects of immigration and their distribution, and try to understand why nation states regulate immigration more restrictively than the mobility of goods and capital. We refer to the example of the free movement of EU citizens. Lastly, we present some economic insights into asylum law.

We wish to thank Jerg Gutmann, Peter Weise and Katharina Eisele for valuable comments and André Plaster for useful research assistance.

Definition

This contribution considers scholarly works on immigration law with a focus on law and economics insights into labor migration and refugee law.

Introduction

Immigration, i.e., the movement of people – usually for permanent residence – into another country or region to which they are not native, is in many respects regulated by the countries concerned. In the following, we discuss some typical motives of migrants (why migrate?), deal with the most important welfare effects of immigration and their distribution (Some basic welfare effects of migration), and try to understand why nation states regulate immigration more restrictively than the mobility of goods and capital and why international agreements on immigration are less frequent than those on trade and investment (Why and how do modern states regulate immigration?). In this chapter, we also discuss the free movement of people in the European Union as an example for a far-reaching cooperation in this field. Finally, we conclude this entry with some ideas on asylum law from an economic perspective (Asylum laws).

Why Migrate?

A fundamental line of research regarding immigration matters concerns individual's motives to migrate. Migration can be triggered by a variety of reasons. Labor migration is a common form; other migration is family based. Some migrants may be attracted by another state's social welfare system. Others expect benefits from committing crimes or terrorist attacks in the country of destination. Many people are forced to leave their

country of origin because of political or religious persecution or are even victims of human trafficking. Needless to say, migrants are often not purely motivated by one reason but by a number of interrelated reasons.

Early efforts to model the migration decision of workers did not take into account the details of the legal environment. They are based on the human-capital model by Sjaastad (1962), which makes the migration decision dependent on the discounted net revenues and the monetary and non-monetary cost of migration. More recent work puts emphasis on the complexity of migration costs (for details see, e.g., Trachtman 2009): They consist, for example, of the direct cost of changing residence, the burden of bureaucratic procedures, the utility loss from abandoning social contacts in the country of origin as well as the time and effort required to establish new contacts in the country of destination, the time and effort required to learn a new language if necessary and to adapt to a different culture, the additional cost of finding an appropriate school for the children, lower pensions for people who have changed their countries of residence during their economically active period more frequently, and so on. At the same time, the benefits may go beyond the remuneration for labor and extend to all kinds of social benefits such as children's allowances, housing subsidies, unemployment relief, social welfare, and free access to schools and universities.

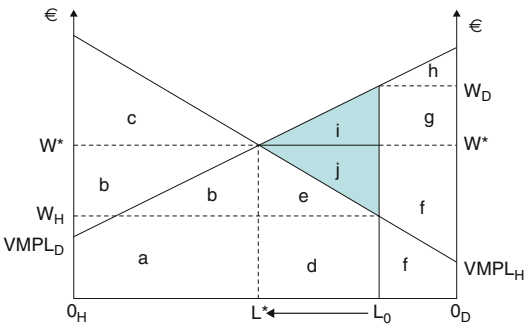
Immigration law affects migration costs by determining who is allowed to enter the country for how long and what are the legal consequences and procedures when people are infringing the law. In a broader sense, immigration law also includes all legal rules that govern access of immigrants to employment, social benefits, and so on.

Some Basic Welfare Effects of Migration

Welfare Effects of Labor Migration

Let us start with a very simple model of the migration of workers. (An excellent presentation of the economic analysis of migrant workers can be found in Borjas (2014). For a recent overview of the economic debate on international migration, see Hatton (2014). Be aware that labor markets are different from goods and capital markets since – different from goods and capital – labor power cannot be separated from the worker. See, e.g., Eger and Weise (1989) with further references.) We assume there are two countries, home country H and country of destination D, there is a given capital stock in each country, capital and goods are completely immobile between the countries, and labor is homogeneous and perfectly mobile between the countries.

Initially, the quantity of homogeneous labor is O_HL_0 in country H and L_0O_D in country D. The curves $VMPL_H$ and $VMPL_D$ represent the value of the marginal product of labor in countries H and D, respectively. With perfect competition on the labor market, the wage rate is equal to the value of the marginal product of labor in each country. Since in the initial situation the wage rate and the marginal product of labor are higher in country D, the reallocation of labor from H to D will produce a gain in allocative efficiency. If there were no migration costs and if migrant workers were induced exclusively by wage differentials, migration would stop only when wages (and the value of the marginal productivity of labor) are the same in both countries. Thus, the total factor income in the receiving country increases (d + e + i + j), while that in the sending country declines by a lower amount (d + e). The welfare gain for both countries together is (i + j). With positive migration costs, the new equilibrium will be somewhere between L* and L_0.



Equilibirum on migration market

However, apart from the overall welfare gain, migration produces even in this simple model winners as well as losers. The winners are migrant labor, gaining (e + j), remaining labor in H, gaining b, and capital owners in D, gaining (g + i), whereas the losers are capital owners in H, losing (b + e), and the existing labor in D, losing g.

Of course, the welfare analysis has to be modified when we relax our strong assumptions. When we allow for heterogeneous labor, it becomes less clear who the winners and losers are. For example, immigrants, who have skills that are complementary to the skill mix of the country of destination, are typically less likely to create losers in this country. On the other hand, the country of origin may suffer losses from emigrating labor force, which used to create positive externalities in that country (so-called brain drain, see, e.g., Sykes 1995; Trachtman 2009; Boeri et al. 2012; Sykes 2013a).

When we allow for inflexible labor markets and unemployment, with homogeneous labor, unemployment will increase in the country of destination and decrease in the country of origin. Without additional assumptions, it is not clear whether unemployment in both countries together will be increasing or decreasing. If immigrants have complementary skills to those of the labor force in the host country, unemployment may decrease in the country of destination (Brücker and Jahn 2011).

When we allow for the mobility of goods, we have to take into account that immigration does not only exert pressure on the wages in the host country but also changes its production and trade structure, for example, toward more labor-intensive production (Hanson and Slaughter 2002). Moreover, it should be mentioned that the mobility of labor is usually accompanied by mobility of capital in the same direction, which is due to the fact that immigration to a country with a fixed stock of capital increases the rate of return on capital (Ottaviano and Peri 2006).

Extending the Basic Model

The provision of labor cannot be looked at in isolation. Based on what we have set out in the previous section regarding individual's reasons to migrate, we have to modify our simple migration model by getting rid of the assumption that migrants are exclusively motivated by differences in gross wages (corresponding to the values of the marginal products of labor). In a more realistic scenario, migrants will also take into account the costs and benefits of the welfare systems in the countries concerned. Actually, migrants will be induced by *differences in net compensation* in the broadest sense, i.e., gross wages minus all kinds of taxes and contributions to the welfare system plus all kinds of social benefits. In this case, part of the migration could be induced by redistributive motives instead of differences in productivity, with at least two possible negative consequences:

- First of all, *people may be induced to emigrate even though their marginal productivity (and their labor income)* is lower in the country of destination than in the country of origin, if the difference in wages is overcompensated by generous social benefits in the country of destination.
- Secondly, if workers with low skills and low income have strong incentives to migrate to countries with generous *social insurance systems*, these systems may *get under pressure* (Sinn 2003, p. 64).

It depends on the respective immigration law and its enforcement to what extent these consequences will arise. Interestingly, for long periods in history, human migration was not subject to any legal constraints (Trachtman 2009, p. 3). However, we can identify market failures that make legal intervention desirable.

Why and How do Modern States Regulate Immigration?

Preliminary Considerations

There is a strong argument from an allocative efficiency point of view to get rid of all restrictions on economic migration (Chang 1997; Trachtman 2009, p. 33; Sykes 2013a). Some scholars claim that there is, hence, no need for an immigration

policy whatsoever (Hayter 2000). It may be desirable to eliminate all immigration controls.

In order to justify a legal intervention, the economist looks for market failures. Looking at immigration generally, the prevalent view among scholars is that international migration can be accompanied by important nonpecuniary externalities (Sykes 2013a, p. 319). These include, by way of example, the additional burden on the welfare state (welfare migration) or may result from the congestion of certain public facilities, from migration which is motivated by higher returns to crime, the import of infectious diseases, and the possibility that migrants may through voting patterns redistribute resources to themselves. These are, hence, a number of varieties of externalities that require at least some border control but in many cases also some regulation of immigration. From a public choice perspective, there is a simple explanation for the existence of regulating the entry of immigrants. As we have shown in the previous chapter, there are winners and losers of immigration. If the losers are better organized than the winners, the former will lobby for a restrictive immigration policy.

The next question when assessing a legal intervention is the welfare standard one wants to apply. There are two main measures to assess immigration policies: One suggests the use of a global welfare function that weighs the welfare of all persons equally. The second trend focuses on national welfare functions in which only the effects on the host countries are considered, excluding the effects on the immigrants' welfare or the welfare of their home countries (Trebilcock 2003). When it comes to arranging for immigration policy at a global scale, hence, the insight that gains are not symmetrical for the countries makes the success of such efforts more unlikely (Sykes 2013b). Even if welfare may be increased at the international level, migration will lead to winners and losers in different countries but also within the countries. Evaluations, hence, depend upon whose welfare is being looked at.

Instruments of National Immigration Policy

Immigration policies in major destination countries, such as the United States, Canada, and Australia, are characterized by a large degree of centralization (see for the following in detail Trebilcock 2003). State authorities conduct basic health, criminal record, and national security checks and rely on a quota system, which distinguishes between three primary classes of immigrants: independent applicants, family members, as well as refugees and asylum seekers. For each class and respective subclasses (e.g., workers with different skills), the states determine quotas limiting the number of respective immigrants. Moreover, the governments issue short-term visas for tourists, students, and temporary workers, who might become immigrants in the future. There are two major problems with this centralized approach of regulating immigration:

1. It is not guaranteed that only those applicants are excluded who are expected to create an unacceptable burden on the amenities of the welfare state. When competing for skilled migrants, burdensome procedures can be a large disadvantage. (With respect to the European Union, see Kocharov 2011.)
2. Since the quotas have to be set in advance, the centralized bureaucracy is confronted with the very difficult (if not impossible) task of predicting the future needs of the labor market.

A further question in this regard is the institutional design by which quotas should be implemented. The basic options are ex ante or ex post screening (Cox and Posner 2007). An ex post system evaluating post-entry conduct is said to provide more information and, hence, a more accurate screening than an ex ante system on the basis of pre-entry information. Overall, both have a number of strengths and weaknesses. A main concern with the ex post system is that immigrants live with the fear of deportation, an uncertainty, which may hamper their integration process to the detriment of the host country.

An alternative approach, which could avoid many of the problems regarding independent applicants and family members and which is to a large degree implemented in the European Union, relies on decentralized agreements between the parties concerned, such as between migrant

workers and employers. In the European Union, various regimes apply. Crucially, migrants' rights differ as to whether the migrant is a European national changing his place of residence from one European member state to another (see International cooperation on migration) or a so-called third-country national entering EU territory. The positions of the latter, furthermore, differ considerably depending on the country that they come from (Eisele 2014). The European Union's major goal is attracting highly skilled workers from non-European countries. (One policy instrument is the Council Directive 2009/50/EC of 25 May 2009 on the conditions of entry and residence of third-country nationals for the purposes of highly qualified employment ("Blue Card Directive").) Whereas the rights of EU nationals are generally aligned, newly acceding states to the European Union may face different types of immigration policies by the various EU member states during a transition period.

A more decentralized system is capable of discouraging irregular immigration. A decentralized system would induce welfare-enhancing and deter welfare-reducing migration, provided there are safeguards that undermine migrants' possibilities to externalize costs to the sending or the receiving country. With respect to the receiving country, the risk of negative externalities can be reduced by (still) centralized health, criminal record, and security checks as well as by a mandatory insurance which avoids that the immigrants become a burden to the welfare state (Trebilcock 2003, p. 296). It is suggested that opening the border would be desirable if at the same time countries made their social programs inaccessible to non-citizens (Sykes 1995). (In Germany, the Academic Advisory Board at the Federal Ministry of Finance made in 2001 the proposal of "delayed integration," i.e., immigrants should enjoy taxed financed social benefits for a transition period from their country of origin. See Sinn (2003, pp. 80–81).) With respect to the sending countries, the problem of "brain drain" may arise if skilled workers, who generate positive externalities in their countries of origin, emigrate. However, one has to take into account that there may be also benefits to the sending countries, such as monetary remittances and positive externalities by those migrants, who improve their skills in the host country and return to the country of origin later on (Trebilcock 2003, p. 282; Boeri et al. 2012).

International Cooperation on Migration

There is no doubt that there are many more barriers to international migration than to international trade and investment. Whereas many barriers to trade have been removed by multilateral, regional, and bilateral agreements and foreign direct investment has been fostered by more than 3,000 bilateral investment treaties (BITs), a comparable degree of international cooperation with respect to migration is (still) missing (Hatton 2007; Trachtman 2009; Gordon 2010). What are the reasons for this mismatch?

There is broad consensus that there are considerable gains from freeing up international migration, even though the estimates differ a lot (Rodrik 2002; Hatton 2007, pp. 345–346). For three reasons, non-coordinated national immigration policies will deviate from the globally efficient policy and will typically lead to over-restrictive immigration policies (Sykes 2013a, p. 320): (1) *Terms-of-trade externalities*: A large receiving country, facing an upward-sloping curve of immigrant labor, may have an incentive to restrict immigration in order to lower the price for immigrant labor. As a consequence, foreign workers absorb some of the costs of the migration restriction. (2) *Enforcement externalities*: Sending countries may have no incentive to reveal information on individuals with contagious diseases, with a propensity to commit serious crimes, or with a poor employment record to the receiving countries, even though they have better access to this information. Without cooperation, all receiving countries face high enforcement cost and will respond with restrictive immigration rules. (3) *Externalities among receiving countries*: Since the immigration policy of one country affects the flow of immigrants to neighboring countries, uncoordinated immigration policy will lead to suboptimal results.

However, to induce international cooperation on immigration, it is not sufficient that gains from cooperation exist. Two other requirements have to

be met to induce cooperation in immigration policy: (1) Gains have to be distributed in a way that all cooperating countries or, more precisely, the relevant interest groups in these countries win and (2) cooperation must be self-enforcing, i.e., all parties should expect that deviations from the cooperative agreement will trigger sanctions by the others. Thus, one has to take into account potential obstacles to beneficial cooperation (Sykes 2013a, p. 327): (1) *The one-way problem*: Different from trade, where typically most countries are interested in exporting goods to and importing goods from the other countries, migration is typically a one-way-issue. People migrate from poor countries to rich countries. Thus, negotiating mutually beneficial agreements becomes more complex, since the parties concerned have to rely on "issue linkage," i.e., in exchange for accepting immigrants from less developed countries, the developed countries have to be compensated in "another currency," e.g., by granting their exporters or investors access to the markets in the less developed countries. (2) *Migration diversion*: In analogy to trade diversion, bilateral or regional agreements on immigration policy discriminate against immigrants from third countries, which may lower social welfare in the immigration countries. One could argue that migration diversion can be avoided if international cooperation adheres to an equivalent to the most-favored-nations obligations under GATT. However, since migration is not controlled by tariffs but by a variety of complex rules and regulations, this obligation would be much more difficult to enforce.

An Example of Far-Reaching Cooperation: The Free Movement of Persons in the European Union

The European concept of the internal market, as defined in the Treaty, is founded on the four fundamental freedoms: the free movement of goods, services, persons, and capital. The underlying idea is to overcome the fragmented goods and factor markets that used to be a characteristic element of Europe after World War II. The free movement of persons between the member states was originally restricted in several ways: (1) The right of free movement was restricted to *nationals* of the member states. (2) The right of free movement was restricted to the *economically active population*, i.e., to workers, self-employed persons, as well as providers of services. (3) Free movement was interpreted as a *prohibition of (direct or indirect) discrimination on the grounds of nationality*. From the very beginning, the Treaty has included a number of explicit derogations from the free movement of persons (public policy, public security, and public health derogations) and a public service exception. In the last 50 years, this situation has been changed by secondary legislation, i.e., in particular directives and regulations and by clarifying judgments of the Court of Justice of the European Union: (1) The *link* between the right of free movement and *economic activity* has been *removed*, by granting this right also to tourists, students, and others. With the Treaty of Maastricht (1993), the decoupling of free movement from economic activity culminated in the recognition of the status of "citizen of the Union" for all nationals of the member states (now: Article 21 TFEU). (2) The right of free movement was extended to *family members who do not have the nationality of a member state*. (3) The prohibition of discrimination on the grounds of nationality was extended to a general *prohibition of obstacles* to the free movement of persons (Brücker and Eger 2012, p. 146).

The Citizens' Rights Directive 2004/38 confirms and extends the right of free movement of European citizens in the following way:

- For *stays of less than 3 months*, the only requirement on Union citizens is that they possess a valid identity document or passport.
- The right of residence for *more than 3 months* remains *subject to certain conditions*: Applicants must be either engaged in an *economic activity* or they must have *sufficient resources and sickness insurance* to ensure that they do not become a burden on the social services of the host member state during their stay.
- *After a five-year period of uninterrupted legal residence*, Union citizens acquire the *right of permanent residence* in the host member state.

Even though the European Union has been removing obstacles to permanent migration between member states to an unprecedented extent, the level of internal migration in the European Union has been modest (see the numbers in Brücker and Eger 2012, p. 158). The reason is that the EU member states, in particular, the "old" 15 members before the accession of Eastern European countries starting in 2004, are characterized by relatively small differences in per capita income levels. On the other hand, language barriers and to some extent also cultural barriers determine migration costs that cannot be removed by legal reforms. But still, there is sufficient evidence that the removal of barriers to the free movement of persons triggered migration, which has contributed to a small but visible increase in the aggregate GDP in the entire European Union (Brücker et al. 2009). Finally, there is no evidence so far that the removal of impediments to the free movements of workers would have triggered a mass inflow of unskilled workers from the East to the West and would have led to the exploitation of the welfare state in the destination countries (Boeri and Monti 2009; Brücker et al. 2009).

Asylum Laws

A rather new field in the economic analysis of the law concerns asylum/refugee law. In order to grant asylum to an asylum seeker, it needs to be determined that he or she is in fact a refugee. The core international agreement in this regard is the 1951 Convention Relating to the Status of Refugees as modified by the 1967 UN Protocol. Importantly, it provides a general definition of "refugee" and stipulates in the so-called non-refoulement clause that a person cannot be forcibly returned to a territory where she may face the risk of persecution. Most countries (and all developed countries) are parties to this Convention. From an economic perspective, international cooperation in asylum matters can be viewed as an agreement among states to supply the global public good of refugee protection (Bubb et al. 2011). While the Convention sets some common ground, national asylum laws also show differences. Lately, one of the key challenges faced by asylum policies is the need to distinguish between "real" refugees and economic migrants. In economic terms, states face the problem of asymmetric information when assessing an individual's status. They have a screening problem. Gradually, some countries' asylum policies have become stricter in an attempt to cope with the challenge of identifying the type of migrant. An economic effect of differing standards in various jurisdictions is that states with stricter policies regarding their admission criteria impose an externality on those countries that have more lenient policies. This, in fact, may induce all countries to raise their standards – a "race to the bottom" (Bubb et al. 2011; Monheim-Helstroffera and Obidzinskib 2010). It is being discussed if deeper integration can alleviate some of the problems resulting from differing standards. On the other hand, transfer systems, according to which wealthy states pay poor states to resettle refugees from other poor states, could create positive externalities on third countries (Bubb et al. 2011).

With a view to the effects of asylum policies, a recent empirical study on the asylum policies in developed countries finds that a tougher asylum policy – regarding entry requirements and conditions in the receiving country – had a deterrent effect on asylum applications (Hatton 2009). It accounted, however, only for one third of the decrease. In looking more closely at the provisions which deter potential asylum seeker, the author finds that the effect is due to those policies that limit access to territory and those that reduce the proportion of successful claims. Provisions, on the other hand, that diminish the socioeconomic conditions of refugees show to have only a low deterrent effect.

Research is also carried out from the perspective of the asylum seeker identifying which migration channel he or she would opt for (Djajić 2014). According to Djajić's model, under current laws, relatively young, skilled, and wealthy asylum seekers, who have access to credit from the family network, are found to have a strong incentive to choose to enter a country with the aid of human smugglers without proper documentation.

Under the Common European Asylum System, Europe has set up a general legal framework for its asylum policy, gradually lifting up more competences to the EU level. In trying to determine the optimal degree of harmonization of Europe's asylum laws, Monheim-Helstroffera and Obidzinskib (2010) develop a regulatory competition model. With respect to the European Union, the authors take a critical stance on ongoing harmonization efforts: In a regulatory context like the European Union, those jurisdictions closest to the external border (peripheral jurisdictions) have an incentive to choose strict admission criteria for refugees. These states would lose most from harmonization efforts that would decrease their legislative discretion. For an EU-wide policy, the question is whether the benefits from harmonization outweigh the losses of the peripheral jurisdiction. From the refugees' point of view, flexibility is warranted as it increases their chances of being admitted to the European Union. While illustrated for the EU context, the model is also more generally applicable. The fear of a "race to the bottom" in the European Union was empirically rejected as national differences in application procedures continued to exist (up to the year 2010 Toshkov and de Haan 2013). The Common European Asylum System has recently undergone some changes; various directives and regulations have been revised and will enter into force in July 2015. Bottom line is that the procedure in the member states have been more aligned, e.g., procedures to apply for asylum have been approximated by the revised Asylum Procedures Directive (Directive 2013/32/EU 2013). In the revised Reception Conditions Directive, common rules have been adopted on the issue of detention of asylum seekers. (Directive 2013/33/EU of the European Parliament and of the Council of 26 June 2013 laying down standards for the reception of applicants for international protection OJ L 180, 29.6.2013, pp. 96–116.) The revised Dublin Regulation seeks to improve the system for determining the member states responsible for the examination of the asylum application by accelerating procedures and reiterating on the decisive criteria, such as recent possession of visa or residence permit in a member state and regular or irregular entry to the European Union. (Commission Implementing Regulation (EU) No 118/2014 of 30 January 2014 amending Regulation (EC) No 1560/2003 laying down detailed rules for the application of Council Regulation (EC) No 343/2003 establishing the criteria and mechanisms for determining the member state responsible for examining an asylum application lodged in one of the member states by a third-country national OJ L 039, 8.02.14, pp. 1–43.)

Conclusions

In this entry, we provided some insights into relevant law and economics contributions on immigration laws and policies. In the light of different forms of migration, a focus on labor/economic migration seemed appropriate, as well as a short excerpt on research topics within asylum law. The relevant contributions illustrate how different types of migrants necessarily impact countries' welfare functions differently. Immigration law is a topic, which can never be a national matter only. Hence, cooperation is a must. We have alluded to some challenges to this cooperation on regional (European Union) or even international level, resulting primarily from potential negative externalities. Legal reforms in immigration law are ongoing, and the topic stimulates a lot of potential future research.

References

Boeri T, Monti P (2009) The impact of labor mobility on public finances and social cohesion. IAB-Deliverable 5

Boeri T, Brücker H, Docquier F, Rapoport H (eds) (2012) Brain drain and brain gain. The global competition to attract high-skilled migrants. Oxford University Press, Oxford

Borjas GJ (2014) Immigration economics. Harvard University Press, Cambridge/Mass

Brücker H, Eger T (2012) The law and economics of the free movement of persons in the European Union. In: Eger T, Schäfer H-B (eds) Research handbook on the economics of European Union law. Edward Elgar, Cheltenham, pp 146–179

Brücker H, Jahn EH (2011) Migration and wage-setting — reassessing the labor market effects of migration. Scand J Econ 113(2):286–317

Brücker H, Baas T, Beleva I, Bertoli S, Boeri T, Damelang A, Duval L, Hauptmann A, Fihel A, Huber P, Iara A, Ivlevs A, Jahn EJ, Kaczmarczyk P, Landesmann ME, Mackiewicz-Lyziak J, Makovec M, Monti P, Nowotny K, Okolski M, Richter S, Upward R, Vidovic H, Wolf K, Wolfeil N, Wright P, Zaiga K, Zylicz A (2009) Labor mobility within the EU in the context of enlargement and the functioning of the transitional arrangements: final report. IAB-final report

Bubb R, Kremer M, Levine DI (2011) The economics of international refugee law. J Legal Stud 40:372–373

Chang HF (1997) Liberalized immigration as free trade: economic welfare and the optimal immigration policy. Univ Penn Law Rev 145(5):1147–1244

Cox AB, Posner EA (2007) The second-order structure of immigration law. Stanford Law Rev 59(4):809–856

Directive 2013/32/EU (2013) Directive 2013/32/EU of the European Parliament and of the Council of 26 June 2013 on common procedures for granting and withdrawing international protection Off J Eur Union L 180:60–95

Djajić S (2014) Asylum seeking and irregular migration. Int Rev Law Econ 39:83–95

Eger T, Weise P (1989) Participation and codetermination in a perfect and an imperfect world. In: Nutzinger HG, Backhaus J (eds) Codetermination. Springer, Berlin/Heidelberg/New York

Eisele K (2014) The external dimension of the EU's migration policy: different legal positions of third-country nationals in the EU: a comparative perspective. Koninklijke Brill, Leiden

Gordon J (2010) People are not bananas: how immigration differs from trade. Northwestern Univ Law Rev 104:1109–1145

Hanson GH, Slaughter MJ (2002) Labor market adjustment in open economies: evidence from U.S. states. J Int Econ 57:3–29

Hatton TJ (2007) Should we have a WTO for international migration? Econ Policy 22:339–383

Hatton TJ (2009) The rise and fall of asylum: what happened and why? Econ J 119:183–213

Hatton TJ (2014) The economics of international migration: a short history of the debate. Labor Econ. https://doi.org/10.1016/j.labeco.2014.06.006

Hayter T (2000) Open borders: the case against immigration controls. Pluto Press, London

Kocharov A (2011) Regulation that defies gravity – policy, economics and law of legal immigration in Europe. Eur J Legal Stud 4(2):9–43

Monheim-Helstroffera J, Obidzinskib M (2010) Optimal discretion in asylum lawmaking. Int Rev Law Econ 30:86–97

Ottaviano GIP, Peri G (2006) Rethinking the effects of immigration on wages. NBER working paper No. 12497

Rodrik D (2002) Final Remarks. In: Boeri T, Hanson G, McCormick B (eds) Immigration policy and the welfare system. Oxford University Press, Oxford, pp 314–317

Sinn HW (2003) The new systems competition. Blackwell, Oxford

Sjaastad LA (1962) The costs and returns of human migration. J Polit Econ 70(4 Suppl):80–93

Sykes AO (1995) The welfare economics of immigration law: a theoretical survey with an analysis of U.S. policy. In: Schwartz WF (ed) Justice in immigration. Cambridge University Press, Cambridge, pp 158–200

Sykes AO (2013a) International cooperation on migration: theory and practice. Univ Chicago Law Rev 80:315–339

Sykes AO (2013b) The inaugural Robert A. Kindler professorship of law lecture: when is international law useful. Int Law Polit 45:787–814

Toshkov D, de Haan L (2013) The Europeanization of asylum policy: an assessment of the EU impact on asylum applications and recognitions rates. J Eur Public Policy 20(5):661–683

Trachtman JP (2009) The international law of economic migration: toward the fourth freedom. W. E. Upjohn Inst. for Employment Research, Kalamazoo

Trebilcock MJ (2003) The law and economics of immigration policy. Am Law Econ Rev 5:271–317

Impact Assessment

Marie-Helen Maras
John Jay College of Criminal Justice, City University of New York, New York, NY, USA

Abstract

Impact assessments help improve the quality of proposed legislation. It enables European Community institutions to determine the economic, social, and/or environmental costs of proposed legislation. It further identifies available options to achieve the objectives of the proposed measure and the positive and negative consequences of each of the options.

Definition

An impact assessment is a tool used by European Community institutions to identify the main policy options available to achieve the objectives of proposed legislation and the intended and unintended costs and benefits of each option. In 2003, the Interinstitutional Agreement on Better Lawmaking set forth basic principles and common objectives for the European Commission, the

European Parliament, and the Council of the European Union for their cooperation during the legislative process and nonlegislative proposals and evaluation of these initiatives (European Parliament, Council and Commission, 2003). In the interest of better lawmaking, decision-makers must have knowledge of the potential economic, social, and environmental impacts (e.g., the internal market, competition, businesses, consumers, employment, human rights, public health, species, or habitats) of initiatives (European Commission, 2009). Since 2003, the European Commission conducted impact assessments to determine this (European Commission, 2006). Specifically, the European Commission conducts impact assessments to determine the economic, social, and environmental consequences of proposed major policy initiatives, including those presented in its Commission's Legislative and Work Programme (or its Annual Policy strategy) and/or nonlegislative proposals with potential significant costs (European Parliament, Council and Commission, 2005). Impact assessments identify the main policy options that are available to achieve the objectives of the proposed initiative, and all relevant stakeholders are consulted during the process. These assessments also provide the costs and benefits of a proposed initiative. The intended and unintended impacts of a proposed initiative are evaluated in quantitative, qualitative, and monetary terms; if the quantification of the consequences cannot be completed, an explanation should be provided in the impact assessment.

The cost-benefit analysis seeks to ensure efficiency in the allocation of resources among competing objectives and proposed actions. What is determined in the impact assessment is whether each policy option represents a more efficient allocation of resources (on a cost-benefit basis) than if such resources were put to alternative use, that is, another policy option. Based on these identified costs and benefits, decision-makers can make informed judgments about proposed initiatives by identifying competing policy objectives and choosing best courses of action. For certain initiatives with potential far-reaching consequences, an extended impact assessment is conducted to provide a more in-depth analysis of the potential impacts of proposals on society, the economy, and the environment.

The impact assessment helps improve the quality of proposals and ensure that they are evidence-based. This process thus assists political decision-making and ensures a comprehensive and transparent approach to the development of legislative and nonlegislative proposals. The Impact Assessment Board, which was created in 2006, is responsible for ensuring the quality of impact assessments (European Commission, 2006). To do so, it evaluates drafts of the European Commission's impact assessments. Once an opinion is issued by the Impact Assessment Board (replaced by the Regulatory Scrutiny Board on July 1, 2015), it accompanies the impact assessment during the European Commission's decision-making on the relevant proposed initiative (European Commission, 2015). The impact assessment is then considered by the European Parliament and the Council of the European Union when evaluating the European Commission's proposals. Pursuant to the 2003 Interinstitutional Agreement on Better Lawmaking, if the European Parliament and the Council of the European Union seek to amend the European Commission's proposal in a substantial manner, they must also conduct an impact assessment, using the European Commission's original impact assessment as a starting point.

References

European Commission (2006) A strategic review of better regulation in the European Union. Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, Brussels. COM 689 final. http://ec.europa.eu/transparency/regdoc/rep/1/2006/EN/1-2006-689-EN-F1-1.Pdf

European Commission (2009) Part III: annex to impact assessment guidelines. http://ec.europa.eu/smart-regulation/impact/commission_guidelines/docs/iag_2009_annex_en.pdf

European Commission (2015) Decision of the president of the European Commission on the establishment of an independent Regulatory Scrutiny Board. C 3263 final. http://ec.europa.eu/smart-regulation/better_regulation/documents/c_2015_3263_en.pdf

European Parliament, Council and Commission (2003) Interinstitutional agreement on better lawmaking. OJ C 321. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV:l10116

European Parliament, Council and Commission (2005) Inter-institutional common approach to impact assessment. http://ec.europa.eu/smart-regulation/impact/ia_in_other/docs/ii_common_approach_to_ia_en.pdf

Impracticability

Hüseyin Can Aksoy
Faculty of Law, Bilkent University, Bilkent, Ankara, Turkey

Abstract

The principle of *pacta sund servanda* requires that agreements must be kept. However such rule is not absolute. When performance of a contractual obligation becomes impracticable, i.e., considerably more burdensome (expensive) than originally contemplated –albeit physically possible- due to an unexpected event, this would lead to adaptation of the contract to the changed circumstances or to avoidance of the contract. In the law and economics literature, impracticability has been substantially studied to figure out who should bear the risk of impracticability; and what would be the efficient remedy for such breach of contract.

Synonyms

Change of circumstances; Economic impossibility; Hardship; Imprévision; Lapse of the contract basis; Wegfall der Geschäftsgrundlage

Definition

An unforeseeable change in the circumstances, arising after the formation of the contract due to an external event, which renders the performance considerably more burdensome albeit physically possible.

Introduction

In the mid-1970s, Westinghouse Electric Corporation sold nuclear reactors to electric companies and also agreed to supply uranium at a fixed price of $8–10 per pound. However, while the company was still obliged to deliver around 70 million pounds of uranium to 27 different buyers, the market price of uranium increased to over $30 per pound. On September 8, 1975, Westinghouse Corporation claimed that performance of its contractual obligation would result in a loss of $2 billion and announced that it would not honor fixed price contracts to deliver uranium. Could the Westinghouse Corporation rely on such change in the circumstances and refrain from performing its contractual obligations?

The principle of *pacta sunt servanda* requires that agreements must be kept. Accordingly, non-performance of any contractual obligation constitutes breach and may cause liability of the debtor. However such rule is not absolute. For instance, in almost all legal systems, it is a well-established principle that when performance is rendered impossible by an event of *force majeure* (e.g., war, earthquake, hurricane, etc.), the debtor will not only be released from his obligation to perform, but this will also eliminate his/her liability to compensate the damages of the creditor. However, the issue is more controversial when performance of a contractual obligation becomes impracticable, i.e., considerably more burdensome (expensive) than originally contemplated – albeit physically possible – due to an unexpected event.

On a theoretical basis, the result of impracticability might be one of the following:

(i) The debtor can be expected to perform despite the increase in costs of performance. Correspondingly, if the debtor fails to perform, he/she has to compensate the damages of the creditor. In such cases, the scope of the recoverable damages (reliance damages or expectation damages) of the creditor must also be determined.
(ii) Impracticability can be regarded as an event, which excuses the debtor. In such cases, the

creditor can neither ask for performance nor compensation. In return, the creditor does not perform either. However it must still be decided if impracticability excuses the debtor *ipso facto* or it grants the debtor a right to avoid the contract.

(iii) It can be argued that impracticability does not lead to an absolute excuse or the avoidance of the contract but only to the adaptation of the terms of the contract to the changed circumstances. In other words, either the parties themselves or the court adapts the contract to the changed circumstances and reestablishes the balance between the debtor's obligation and the creditor's counter-obligation.

Impracticability in Modern Legal Doctrine

The modern legal doctrine regards impracticability as an exception to the principle of *pacta sunt servanda* provided that the performance becomes excessively difficult for the debtor due to an external event, which could not be foreseen at the time of the conclusion of the contract, and under the new circumstances the debtor cannot reasonably be expected to perform as stipulated in the contract.

It must be emphasized that the border between impossibility and impracticability is hard to draw. Despite their similarities, these two concepts have important differences as well. Since impossibility is regarded as an objective and permanent obstacle to performance, it causes expiration of the primary obligation of the debtor. This follows from the fact that the debtor cannot be expected to perform what is impossible. On the other hand, in case of impracticability the obligation is – at least theoretically – still possible, however at enormous and unexpectedly high cost. Therefore, in such cases, the primary tendency is to preserve the contract, and it is generally ruled that the contractual balance must be restored through adaptation of the obligations of the parties to the changed circumstances. However, if such adaptation is not possible, the second option is the avoidance of the contract. This would result in elimination of the obligations of both parties, including compensation liability.

When performance becomes more burdensome – albeit possible – some legal orders, for instance, the German Civil Code, distinguish between two case groups, which lead to different legal consequences. According to §275 II BGB, if there is gross disproportionality between the creditor's interest in performance and the debtor's interest in non-performance (i.e., required efforts and expenses to perform), the debtor may refuse to perform. The classical example of this case is the ring which falls into the sea. In this case, the legal consequence of impracticability is the same as that of impossibility: the debtor's obligation to perform the primary obligation extinguishes. A second group captures those cases, in which the cost of performance rises enormously, just as in the first group, after the conclusion of the contract, but the value of the contract (performance) rises too. In such cases, which fall under §313 BGB, there is no disproportionality, but the circumstances which became the basis of the contract significantly change. For example, an art dealer buys a painting of a famous painter from a gallery which the gallery must still buy on the market. But before the purchase by the gallery, a fire in a museum destroys many paintings of the particular painter, which increases the price of the bought painting by the factor 10. Here the event which makes the performance so expensive increases proportionately the interest of the creditor in performance. In this case, the debtor's obligation to perform the primary obligation would not extinguish but the legal consequence would be adaptation of the contract by raising the price. Revocation would be the secondary option, if adaptation is not possible or one party cannot reasonably be expected to accept adaptation.

The practical difference between the scopes of these provisions is the following: when the value of the performance increases, the creditor's interest in receiving performance increases proportionally as well; hence the performance which has become burdensome for the debtor is not disproportionate to the interest of the creditor. Therefore, an objective increase in the market price of the good is separated from the cases, where the

procurement of the good has become particularly costly for the specific debtor.

Economic Analysis of Impracticability

Impracticability has been substantially studied in the law and economics literature. In fact, economic considerations might play an important role in drawing the line between a bearable difficulty and "excessive difficulty" (impracticability). Moreover, economic reasoning might be useful in determining when the parties can resort to avoidance of the contract instead of adaptation to the changed circumstances.

Risk-Bearing Perspective

In the earlier stages, authors approached the issue from the economic theory of efficient risk bearing and imposed the entire results of impracticability on either of the parties (Posner and Rosenfield 1977; Joskow 1977). Within this stance, in their pioneering study, Posner and Rosenfield argue that a discharge question arises only if the contract has not assigned the risk in question (to either of the parties) and the event, giving rise to the discharge claim was not avoidable by any cost-justified precautions (of the debtor). Provided that these two conditions are met, the authors argue that the loss should be placed on the party who is the superior (the lower-cost) risk bearer. Accordingly, the superior risk bearer can be figured out by three factors: knowledge of the magnitude of the loss, knowledge of the probability that it will occur, and costs of self-insurance or market insurance. For instance, the party, who is in the position to prevent the materialization of the risk or to insure the risk at a lower cost, would be the superior risk bearer. In the end, if the debtor is found to be the superior risk bearer, he/she must perform despite the increase in the costs of performance. However should the creditor be in the position to bear the risk, the debtor would be discharged.

Efficient Remedy Perspective

Posner and Rosenfield's all-or-nothing approach, which shifted the entire risk to either of the parties, was later on questioned by some authors, who focused their studies on designing efficient remedies for breach of contract. Instead of deciding who should bear the entire risk, these authors associated impracticability with efficiency and focused on dividing the risk between the contracting parties. Within this context, considering the efficiency of the remedy in a given case, the result of impracticability could vary between expectation damages and no remedy.

With efficiency considerations, it is generally argued that except for some very rare cases, discharge of the debtor would not yield to efficient results (White 1988; Sykes 1990). As a matter of fact, associating impracticability with discharge of a contract will negatively affect the risk bearing of the parties. Moreover the availability of discharge as a remedy will create high breach incentives on the debtor.

Even if it is accepted that impracticability should not lead to discharge of the debtor except for exceptional cases, it is argued that neither expectation damages nor reliance damages must be covered in all cases of impracticability. In fact such decision must be rendered on a case-by-case analysis. Within this stance, some authors propose to focus on the risk aversion of the parties while making a choice between the imposition of expectation damages and reliance damages (Shavell 1980; Polinsky 1983) or even between expectation damages and no damages (Sykes 1990). Others argue that the parties' expectancies at the time of the conclusion of the contract regarding the change of circumstances will be decisive on the choice between expectation and reliance damages (Eisenberg 2009).

Another group of authors argue that in cases of impracticability, the appropriate remedy would be adjustment of the contractually contemplated price (Speidel 1981; Trakman 1985; Trimarchi 1991). The main arguments of these authors is that in cases of absolute uncertainty (in systematic risks such as inflation or international crises, which affect the economy as a whole), the "superior risk bearer" criterion is inappropriate and insurance is not an adequate option (Trimarchi 1991). Moreover, it is asserted that the price adjustment of the court redresses the advantaged party's opportunistic conduct as the advantaged

party has the duty to cooperate and bargain in good faith ex-post (Speidel 1981).

Another proposal is that the total surplus of the contract must be taken into consideration, when deciding on the result of impracticability (Aksoy and Schäfer 2012). Within this stance, following the increase in costs of performance, if the contract still generates positive surplus, there is no reason to avoid the contract. In this case, if the risk was very remote and both parties failed to consider the risk, the obligations of the parties can be adapted to the changed circumstances. However if the total surplus is obviously highly negative and if this is observable by a third party like the judge, the contract would be avoided. In this case, in addition to the excessive and low probability cost increase, the increase in relation to the interest of the buyer (consumer surplus) triggers avoidance of the contract.

Finally, it must be emphasized that the necessity of the excuse doctrines is disputed itself. For instance, there are some authors who question the well-recognized assumption that parties cannot allocate risks under uncertainty. According to Triantis (1992), the question is not whether a particular risk is allocated or not, but at what level it is allocated. The author argues that even if all risks are not "explicitly" allocated, they can be allocated under broader risk groups. For instance, an increase in the transportation costs due to increase in oil prices following a nuclear accident in the Middle East cannot be "explicitly" anticipated, but the parties may allocate the broader risk of a large increase in oil prices for any reason. Therefore, allocation of risks must be left to the parties and the role of the contract law should be restricted to interpretation and enforcement of the risk allocations of the parties.

References

Aksoy HC, Schäfer HB (2012) Economic impossibility in Turkish contract law from the perspective of law and economics. Eur J Law Econ 34:105–126

Eisenberg MA (2009) Impossibility, impracticability, and frustration. J Legal Anal 1:207–261

Joskow PL (1977) Commercial impossibility: the uranium market and the Westinghouse case. J Legal Stud 6:119–176

Polinsky AM (1983) Risk sharing through breach of contract remedies. J Legal Stud 12:427–444

Posner R, Rosenfield A (1977) Impossibility and related doctrines in contract law: an economic analysis. J Legal Stud 6:83–118

Shavell S (1980) Damage measures for breach of contract. Bell J Econ 11:466–490

Speidel RE (1981) Court-imposed price adjustments under long-term supply contracts. Northwest Univ Law Rev 76:369–422

Sykes AO (1990) The doctrine of commercial impracticability in a second best world. J Legal Stud 19:43–94

Trakman LE (1985) Winner take some: loss sharing and commercial impracticability. Minn Law Rev 69:471–519

Triantis GG (1992) Contractual allocations of unknown risks: a critique of the doctrine of commercial impracticability. Univ Toronto Law J 42:450–483

Trimarchi P (1991) Commercial impracticability in contract law: an economic analysis. Int Rev Law Econ 11:63–82

White MJ (1988) Contract breach and contract discharge due to impossibility: a unified theory. J Legal Stud 17:353–376

Further Reading

Christopher B (1982) An economic analysis of the impossibility doctrine. J Legal Stud 11:311–332

Goldberg V (1985) Price adjustment in long-term contracts. Wisconsin Law Rev 1985:527–543

Gordley J (2004) Impossibility and changed and unforeseen circumstances. Am J Comp Law 52:513–530

Perloff JM (1981) The effects of breaches of forward contracts due to unanticipated price changes. J Legal Stud 10:221–235

Smythe DJ (2011) Impossibility and impracticability. In: De Geest G (ed) Contract law and economics, encyclopedia of law and economics, 2nd edn. Edward Elgar, Cheltenham/Northampton, pp 207–224

Smythe DJ (2004) Bounded rationality, the doctrine of impracticability, and the governance of relational contracts. S Calif Interdisc Law J 13:227–267

Triantis GG (1992) Contractual allocations of unknown risks: a critique of the doctrine of commercial impracticability. Univ Toronto Law J 42:450–483

Wright AJ (2005) Rendered impracticable: behavioral economics and the impracticability doctrine. Cardozo Law Rev 26:2183–2215

Imprévision

► Impracticability

Imprisonment

► Prisons

Incarceration

► Prisons

Incomplete Contracts

Maria Alessandra Rossi
Department of Economics and Statistics, University of Siena, Siena, Italy

Abstract

The notion of incomplete contracts refers to the circumstance that some aspect of contractual parties' payoff-relevant future behavior or some relevant payoff in future contingencies is unspecified in the contract and/or unverifiable by third parties. This may be attributed, by and large, to three different causes: high enforcement costs entailing unverifiability by third parties such as courts or arbitrators; the transaction costs that arise from uncertainty about future events, from the contractual parties' bounded rationality, and from judges' bounded rationality; and, finally, from asymmetric information. Different research programs in the economics of contracting explore the implications of these different sources of contractual incompleteness, providing insights addressing an extremely wide range of contractual issues, including the theory of the firm, the theory of corporate finance, the analysis of formal and informal institutions, regulation and public ownership, innovation and intellectual property, and international trade. The extent to which the notion of contractual incompleteness also has relevant normative implications for the law and economics of contract regulation is an issue currently debated.

Definition

Contracts can be considered incomplete, from an economic perspective, when some aspect of parties' payoff-relevant future behavior or some relevant payoff in future contingencies is unspecified in the contract and/or unverifiable by third parties. In other words, an incomplete contract is a contract that is insufficiently state-contingent, so that some or all of the parties to the transaction are unsure as to the effective payoffs they will receive in any future state of the world. When the cause of incompleteness is attributed entirely to the fact that some aspects of the transaction are observable to the parties but unverifiable by third parties, an "incomplete contract approach" is said to be adopted.

Introduction

The notion of incomplete contract belongs more to the realm of microeconomic theory than to the law and economics approach *stricto sensu*. The scholarly contributions that have explored the wide-ranging implications of the notion have indeed focused mostly on the positive analysis of contractual parties' behavior with regard to the choice of the contractual form that may best limit the negative effects of incompleteness and on the comparative analysis of the choice of the governance arrangement most apt to complement or substitute for the incomplete contract. Thus, most of the literature that attributes relevance to the notion of contractual incompleteness does not directly address the issue of the effects of legal rules on contractual parties' behavior that is the core concern of the law and economics analysis of contracts. In other words, the literature on incomplete contracts deals mostly with the mechanics of private contracting rather than with contract law. There is, nonetheless, some overlap, to the extent that the economic notion of incomplete contract and the associated research programs provide normative implications for contract interpretation and contract regulation.

This entry provides an overview of the concept, considering first its meaning according to the

different theories that have highlighted its relevance and then its many implications in a wide range of research domains. The insights drawn from the notion of incomplete contract have indeed been fruitfully applied to a wide number of specific issues, including the theory of the firm, the theory of corporate finance, the analysis of formal and informal institutions, regulation and public ownership, innovation and intellectual property, and international trade. The normative implications of contractual incompleteness for the law and economics of contract regulation are also briefly reviewed.

The Meaning of Contractual Incompleteness

A contract may be said to be literally incomplete if it contains a true "gap," namely, if it does not contain provisions relating to some event or circumstance that may arise in the future (*unforeseen contingency*) and therefore does not define parties' behavior in such circumstances. While this notion of contractual incompleteness has been adopted, implicitly or explicitly, in a number of law and economics analyses, the expression "incomplete contract" is more commonly understood to refer to the strictly economic version of the concept, which introduced into microeconomic theory the general insight that real-world contracts may significantly diverge from the perfect, fully state-contingent contracts depicted by the theory of perfectly competitive markets.

Thus, incomplete contracts may be defined, from an economic standpoint, as insufficiently state-contingent contracts (see, e.g., Schwartz 1998). The economic literature identifies three possible reasons for their existence, to which correspond, by and large, three research programs in the economics of contracting.

The first reason for contractual incompleteness is given by enforcement costs and is emphasized by the research program inaugurated by Sanford Grossman, Oliver Hart, and John Moore and defined as "incomplete contract theory," "new property rights approach," or GHM theory. According to this perspective, contracts are incomplete because some aspects of the transaction are observable by the parties but not verifiable by a third party in charge of enforcement – a judge or an arbitrator. In other words, the origin of contractual incompleteness should be attributed to a problem of asymmetry of information between the contractual parties and a relevant outsider – a judge or an arbitrator – in charge of enforcing the contract.

The second source of contractual incompleteness – emphasized by new institutional and transaction cost economics – is given by the transaction costs that arise from uncertainty about future events, from the contractual parties' bounded rationality, which limits their ability to account ex ante for all the future contingencies that may affect parties' contractual payoffs, and from judges' bounded rationality. The relevant transaction costs may be of at least four main types: (1) the cost of foreseeing all the possible states of the world that may materialize during the contractual relationship; (2) the cost of negotiating and finding an agreement about the appropriate course of action parties should take in any future state of the world; (3) the cost of describing ex ante in a contract the characteristics of what is traded and/or the parties' effort for each possible contingency; and, finally, (4) enforcement costs (Williamson 1985; Hart 2008). According to this perspective, incompleteness may also be endogenous because parties rationally decide to save on contracting costs when transaction costs exceed the corresponding benefit.

A third cause of contractual incompleteness is asymmetric information between contractual parties, explored by principal-agent theory. This perspective entails that contracts are voluntarily left incomplete because more complete contracts would create opportunities for moral hazard or adverse selection. Making payoffs contingent on future states of world that cannot be symmetrically observed by parties may indeed allow the informed party to misrepresent the state of the world that has materialized so as to increase her payoffs. Similarly, in some instances, a complete contract may not be chosen because it reveals to the counterpart valuable private information. In

all of these cases, incompleteness is entirely endogenous as it emerges from the equilibrium choices of players whose actions are not constrained by bounded rationality or asymmetric information with respect to third parties. This third source of contractual incompleteness is worth mentioning for completeness, but it should be noted that the notion of "incomplete contract" is generally evoked by reference to transaction and/or enforcement costs.

Thus, there is some debate on the root causes of contractual incompleteness. The debate goes as far as to question the very existence of a theoretical foundation for the notion of incomplete contracts and to suggest the superiority of complete contracting approaches, related to implementation theory and mechanism design (Schmitz 2001). The main criticism to the notion of incomplete contracts is that the assumption of unverifiability by third parties appears weak if parties are symmetrically informed (Maskin and Tirole 1999). Payoff-relevant information is generally only partly unverifiable and parties may invest resources to increase verifiability. Moreover, symmetrically informed parties may adopt complex revelation mechanisms that can be negotiated ex ante and that ensure that both parties will have incentives to reveal their true preferences ex post, so as to ensure efficient outcomes. It should be noted, however, that these complex contracts are seldom observed in practice.

The Holdup Problem

Contractual incompleteness matters, from an economic standpoint, in so far as it affects the realization of Pareto-efficient exchanges. Most of the incomplete contracting literature focuses on circumstances when this is the case because the contract involves some form of specific investment and at least an opportunistic party (Williamson 1985).

Investments specific to particular assets or relationships are investments whose economic value is much higher within the relationship and provided that the investing party has access to the relevant assets rather than outside the relationship or in absence of access to the assets. In other words, the ex post value of specific investments outside of the relevant transaction is much lower than their ex ante next best alternatives. Thus, once specific investments have been incurred, the contractual parties become to some extent locked into each other (what Oliver Williamson has famously dubbed the "fundamental transformation" of a standard transaction into a bilateral monopoly).

Agents who make specific investments are therefore exposed to the risk of counterparts' opportunistic behavior because, to realize the surplus from their investments, they need the cooperation of their contractual counterparts and they require access to a particular set of assets. However, precisely because of this *lock in effect*, contractual counterparts may behave opportunistically and try to renegotiate the terms of the original contract so as to obtain a greater share of the total surplus (this is the substance of what is called the "holdup problem"). Since the level of specific investment is not verifiable ex post, the extent to which an individual will be able to appropriate the surplus from his investment cannot be determined ex ante via the original contract and depends on his ex post bargaining power. The source of inefficiency therefore lies in the fact that the threat of holdup constitutes a deterrent to the ex ante realization of unverifiable specific investments, so that some Pareto-efficient transactions may be inhibited.

Incomplete Contracts and the Theory of the Firm

The first and foremost application of the incomplete contracting framework concerns the theory of ownership and vertical integration developed by the incomplete contracts approach on the basis of insights first proposed by Oliver Williamson. This theory explains the very existence and the boundaries of the firm by emphasizing the important economic function of the allocation of property rights over physical or nonhuman assets in a situation of incomplete contractibility (Grossman and Hart 1986; Hart and Moore

1990). When contracts are incomplete, ownership of physical or nonhuman assets is indeed important because it influences parties' threat points in the ex post bargaining and therefore the division of ex post surplus. This is because "the owner of an asset has residual control rights over that asset: the right to decide all usages of the asset in any way not inconsistent with a prior contract, custom or law" (Hart 1995, p.30). Ownership ensures ex ante the owner that he will be able to dispose ex post of the asset and will not be excluded from its use. The increased bargaining power at the renegotiation stage provides the owner with a greater incentive to invest with respect to non-owners.

The incomplete contracts approach thus conceptualizes the firm as a collection of assets over which the owner has residual rights of control and proposes a theory of optimal ownership allocation according to which ownership rights over nonhuman assets should be assigned to the agents who value them the most, i.e., to the parties who have to make the most relevant and specific investments in human capital. Assuming that parties may efficiently bargain ex ante on the allocation of ownership rights, the theory also predicts that efficient ownership allocations will tend to emerge in equilibrium (for a survey of contributions developing this approach, see Aghion and Holden, 2011).

The solution envisaged represents, however, only a second-best solution and the theory has the merit of highlighting also the costs of ownership allocation as a mechanism to align incentives. The most relevant of these costs is that the incentive provided through the allocation of residual control operates only with respect to the owner, while incentive problems persist with regard to non-owners. When the number of agents required to make specific investments is high, the gap between the first-best and the second-best solutions will be particularly wide, and the allocation of ownership rights will therefore display a limited efficacy as an incentive mechanism.

The incomplete contracting approach provides a framework for the comparative analysis of governance structures that may explain vertical integration choices and the scope of the firm's boundaries. The theory is, however, subject to a major criticism: there is some logical tension in assuming that agents may not sign complete contracts specifying required specific investments but that they can perfectly foresee ex post payoffs in order to bargain on the optimal ownership allocation.

The key role played by the allocation of residual rights of control in an incomplete contract framework has been explored by this literature also by a different angle. Rather than focusing on the allocation of ownership as a mechanism to align incentives, a large number of contributions has explored the design of contractual procedures that define the ex post renegotiation framework in a way that, by constraining renegotiation, suitably allocating bargaining power and defining default options allows to overcome the holdup problem (Chung 1991; Aghion et al. 1994). The mechanisms thus defined (*option contracts*) allow to achieve first-best outcomes but require higher degrees of verifiability than GHM-style models because the external enforcer is assumed to be able to recognize who is the opportunistic party (for a survey, see Schmidt 1998). This strand of research seeks to explain contractual behavior rather than firm's organizational choices. The main limit of this approach stressed by critics is that the theory's basic insights are very sensitive to underlying assumptions.

Incomplete Contracts, Transaction Costs, and Institutions

Starting from a somewhat stronger notion of contractual incompleteness than simple observability plus non-verifiability, the new institutional and transaction cost research program focuses on the comparative analysis of the institutional arrangements designed to mitigate the effects of incompleteness (see, e.g., Brousseau and Glachant 2002). Given that agents have limited cognitive resources and face strong (Knightian) uncertainty, according to this perspective, they devise their contractual relationships in ways that allow for the minimization of transaction costs. The solutions that emerge from transaction cost

minimization vary from bilateral contracts to formal and informal institutions, all of which may perform similar functions in terms of ensuring effective enforcement of the contractual relationship. The contractual relationship is conceptualized as a sophisticate object designed to provide safeguards for specific investments, promote parties' commitments, and guarantee performance through the definition of negotiation procedures, supervision mechanisms, and conflict-resolution tools. It implements what this literature calls a "private order" meant to discipline parties' behavior.

The core issue involved by contract design is given by the trade-off between opportunism and efficient adaptation (see, e.g., Nicita and Pagano 2005). Contracts may be long or short, detailed or open-ended. The choice among these features depends on their relative costs and benefits: when contract duration is long, a very detailed contract may be chosen with the purpose of minimizing the risks of opportunism at the cost of incurring higher ex ante transaction costs and lower ex post flexibility. At the same time, the broader and more open-ended the contract, the higher the flexibility allowed to the parties in order to efficiently react and adjust the mechanics of their relationship to unforeseen contingencies. This basic trade-off gives rise to the emergence of contracts with different durations and different degrees of ex ante specification of parties' obligations, according to the nature of the underlying relationship and to the characteristics of the institutional environment.

Institutions play a key role for contractual design because they provide the default rules of the game that parties may choose instead of elaborating more complex bilateral arrangements. Relevant institutions include, of course, public institutions in charge of enforcement such as the legal system and the judiciary but also a wide range of other institutional arrangements that have a private nature, both formal (codes of conduct, business associations, standard-setting organizations, etc.) and informal (customs, reputation, corporate culture, etc.) (North 1990). The nature of the institutional environment influences the choice of contractual terms and the ability of given contracts to implement efficient outcomes.

The logic of transaction cost minimization of the new institutional and transaction cost approach is meant to explain the emergence of different institutional arrangements as efficient responses to given transactional characteristics and features of the institutional environment. However, it does not easily lend itself to explain the persistence of inefficient institutions and private arrangements. The literature on institutional complementarities (Pagano and Rowthorn 1994; Aoki 2001) also starts from the notion of contractual incompleteness but provides an explanation for inefficiencies. It focuses on the existence of multiple equilibria across different choice domains: agents are unable to coordinate their choices across different domains, and therefore, their choices in one domain are influenced by the choices made in other domains, giving rise to a multiplicity of institutional arrangements, each of which constitutes a Nash equilibrium with self-reinforcing properties. Depending on initial conditions, inefficient equilibria may emerge and persist through time, unless exogenous shocks intervene to shift the system toward a different equilibrium.

Incomplete Contracts and Corporate Finance

The incomplete contract notion has shed new light also on the analysis of the determinants of the choice of a firm's financial structure, providing a perspective that highlights for the first time the importance of corporate finance choices for incentives. The incomplete contracts approach to corporate finance allows to overcome some of the limits of the so-called trade-off theory of the optimal capital structure of the firm according to which the optimum debt-equity ratio is chosen by equating the marginal benefit in terms of tax savings from the use of debt (which the tax treatment renders cheaper than equity after tax) with the expected marginal bankruptcy cost. This theory, prominent in corporate finance, does not address the issue of the impact of the capital structure on stakeholders' incentives to make firm-specific investments and therefore on firm performance.

Adopting an incomplete contracts approach allows, by contrast, to trace a link between corporate finance and the generation of a firm's cash flow by analyzing the effect of alternative financial structures on the allocation of control rights. It is when enforcement of both financial contracts and managers' incentive contracts is limited that the allocation of residual rights of control matters (Bolton 2013). This is because financing choices affect the distribution of decision powers in the event of unforeseen contingencies and therefore shape ex ante incentives.

This perspective thus provides an economic justification for the use of debt different from tax advantages: debt is a form of financing that ensures a flexible and contingent allocation of control rights between the investor and the entrepreneur (Aghion and Bolton 1992). Recourse to debt implies that control rests in the hand of the entrepreneur in case a good state of the world materializes and that it shifts in the hands of the financier in case of a bad state of the world. This provides safeguards to both parties: to the financier, who is shielded against excessive losses in the bad state, and to the entrepreneur, who is protected from the risk of losing control in the good state.

Incomplete Contracts and Innovation

Both contractual incompleteness and asset specificity appear especially pronounced in the context of innovative activities. Innovation is indeed a collective, cumulative, and highly uncertain process that involves specific investments by a large and varied number of firm stakeholders (financiers, managers, workers, etc.) who contribute financial, physical, and intellectual resources to a common endeavor whose final outcome is often to a large extent unpredictable and impossible to specify in a detailed contract. The risk of underinvestment associated to the holdup problem is thus magnified in this context, as the collective nature of the innovative process multiplies the instances of potential holdup.

Acknowledgement of these features has prompted a number of contributions exploring the implications of incomplete contracts for firms' ability to innovate. This literature takes as a starting point the basic intuition of the incomplete contract approach, namely, that absent the possibility to write complete contingent contracts, the rules affecting the allocation of residual rights of control over the relevant assets deeply influence stakeholders' incentives to invest. Therefore, this stream of research focuses on the relationship between the institutions, market forces, and internal governance arrangements that jointly define firms' corporate governance structure and innovative performance. In an incomplete contracting framework, corporate governance rules matter because they affect the extent to which financial investors are guaranteed a return for their investments, and therefore the price at which there are willing to provide the firm with the funds necessary to undertake innovative projects, and because they affect the distribution of residual rights of control within the corporation, and therefore stakeholders' incentives to make specific investments in human capital. A rich theoretical and empirical literature has uncovered the many facets of firms' corporate governance that may have a bearing on innovation, including the degree of ownership concentration, owners' identity, firms' capital structure, the extent of workers involvement in firms' decision-making processes, and institutional factors such as the degree of unionization and the rules disciplining takeovers (for a survey, see Belloc 2012). A related strand of research adopts an incomplete contract framework to explore the costs and benefits of the allocation of intellectual property rights as an incentive mechanism (Aghion and Tirole 1994; Pagano and Rossi 2004).

Incomplete Contracts and International Trade

Another domain where it is natural to assume the presence of contractual incompleteness is international trade. International transactions occur in absence of effective enforcement systems, since uncertainties exist as to applicable laws when contracting parties reside in different countries,

existing international conventions and trade fora are limited at best, and implicit contracts based on repeated interactions are scarcely effective in curbing opportunism. The notion of incomplete contracts has indeed been fruitfully applied in open-economy environments to study the issue of firm organization and vertical integration in the international context so as to explain the determinants of multinational activity and the structure of international trade flows. In particular, attention has been devoted to explaining, on the basis of GHM-style arguments, why multinationals tend to integrate capital-intensive productions of intermediate inputs and to outsource labor-intensive productions (Antràs 2003). According to this literature, the choice reflects the allocation of property rights to the party who has to make the most important specific investment.

The other main strand of research exploring the implications of the concept of contractual incompleteness in an open-economy environment aims at explaining countries' comparative advantage in the production of goods requiring relationship-specific investments. Since countries differ in their ability to ensure contract enforcement, and since the effectiveness of enforcement affects incentives to make relationship-specific investments, countries with stronger contract enforcement will have a cost advantage in productions requiring relatively higher degrees of relationship-specific investments. This may explain patterns of international trade as well as firms' geographical location (see, e.g., Nunn 2007).

These two strands of research do not exhaust by any means the landscape of contributions linking incomplete contracts to international trade but nonetheless provide a useful introduction to the nature of the issue addressed. The literature on trade and institutions inspired by the incomplete contracts notion has, indeed, flourished both theoretically and empirically in recent years, providing insights for analyzing trade policy choices, understanding the role of power in international transactions, the financial structure of multinational firms, and many other globally relevant issues (for surveys, see Antràs 2013; Helpman 2006).

Incomplete Contracts and the Law and Economics of Contract Regulation

From a strict law and economics perspective, the economic problem of incomplete contracts matters because it is at the roots of the legal problem of contract regulation, which includes the issue of contract interpretation. Contractual incompleteness implies that there is a role for the State in filling the gaps left by parties by interpreting the contract, supplying a common framework and vocabulary to contracting parties, supplying default rules, and eventually regulating the contracting process through other means. What exactly this role should be has, however, so far not been fully defined by the incomplete contracts literature.

A relevant point of view in this regard holds that the notion of contractual incompleteness should be taken to suggest that the State should refrain from attempting to provide efficient default rules to "complete" contracts because it is highly unlikely to possess more accurate information or to incur lower transaction costs than the parties in drafting the relevant provisions (Hermalin and Katz 1993). Moreover, when contracts are endogenously incomplete, any attempt to provide default rules that modify parties' original contractual arrangement may undermine the latter by modifying the terms of the chosen renegotiation game. A further implication of this line of reasoning is that specific performance should be preferred to any attempt to regulate contracts by supplying default rules or imposing efficient solutions (Schwartz 1998).

A different viewpoint holds that the normative implications of the incomplete contracting literature are limited and inconclusive, mostly because the latter does not appear to successfully predict either contract content or interpret legal doctrines such as the penalty doctrine (Posner 2003).

Conclusion

The notion of incomplete contract has gained wide currency in both microeconomics and law and economics. The concept appears to some extent elusive: while most would agree that

non-verifiability of contractual terms by third parties and insufficient state-contingency of the contract are key elements of the definition, there is some theoretical debate on other relevant aspects and on the very meaning and foundations of the concept. In spite of these debates, however, the concept has proven to be rather versatile, as it underlies, in one form or another, a diverse collection of both theoretical and empirical studies. Starting from the basic application in explaining the size and the boundaries of the firm and in motivating comparative analysis of institutional arrangements, the notion of contractual incompleteness has spurred research in a wide range of more specific domains that has refined our understanding of firms and institutions. The appropriate way in which these insights should be incorporated into the analysis of the effects of legal rules on economic agents' behavior is, however, still somewhat controversial and deserves further study.

Cross-References

- ▶ Governance
- ▶ Institutional Complementarity
- ▶ Institutional Economics
- ▶ Transaction Costs

References

Aghion P, Bolton P (1992) An incomplete contracts approach to financial contracting. Rev Econ Stud 593:473–494

Aghion P, Holden R (2011) Incomplete contracts and the theory of the firm: what have we learned over the past 25 years? J Econ Perspect 252:181–197

Aghion P, Dewatripont M, Rey P (1994) Renegotiation design with unverifiable information. Econometrica 62:257

Aghion P, Tirole J (1994) The Management of Innovation. The Quarterly Journal of Economics Vol. 109(4), pp. 1185–1209

Antràs P (2003) Firms, contracts, and trade structure. Q J Econ 118(4):1375–1418

Antràs P (2013) Goes global: incomplete contracts, property rights, and the international organization of production. J Law Econ Organ. First published online 17 Feb 2013. https://doi.org/10.1093/jleo/ews023

Aoki M (2001) Toward a comparative institutional analysis. MIT Press, Boston

Belloc F (2012) Corporate governance and innovation. J Econ Surv 265:835–864

Bolton P (2013) Corporate finance, incomplete contracts, and corporate control. J Law Econ Organ. First published online 9 Oct 2013. https://doi.org/10.1093/jleo/ewt010

Brousseau E, Glachant J-M (2002) The economics of contracts and the renewal of economics. In: Brousseau E, Glachant J-M (eds) The economics of contracts. Theories and applications. Cambridge University Press, Cambridge

Chung T-Y (1991) Incomplete contracts, specific investments and risk sharing. Rev Econ Stud 58:1031

Grossman SJ, Hart OD (1986) The costs and benefits of ownership: a theory of vertical and lateral integration. J Polit Econ 944:691–719

Hart OD (1995) Firms, contracts, and financial structure. Oxford University Press, Oxford

Hart O (2008) Incomplete contracts. In: Durlauf SN, Blume LE (eds) The new Palgrave dictionary of economics, 2nd edn. Palgrave Macmillan, London

Hart OD, Moore J (1990) Property rights and the nature of the firm. J Polit Econ 98:1119–1158

Helpman E (2006) Trade, FDI, and the organization of firms. J Econ Lit 44(3):589–630

Hermalin BE, Katz ML (1993) Judicial modification of contracts between sophisticated parties: a more complete view of incomplete contracts and their breach. J Law Econ Organ 9:230

Maskin E, Tirole J (1999) Unforeseen contingencies, property rights, and incomplete contracts. Rev Econ Stud 66:83–114

Nicita A, Pagano U (2005) Incomplete contracts and institutions. In: Backhaus A (ed) Elgar companion to law and economics. Edward Elgar, Cheltenham, pp 145–164

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

Nunn N (2007) Relationship-specificity, incomplete contracts and the pattern of trade. Q J Econ 1222:569–600

Pagano U, Rossi MA (2004) Intellectual property rights, incomplete contracts and institutional complementarities. Eur J Law Econ 182:55–76

Pagano U, Rowthorn R (1994) Ownership, technology and institutional stability. Struct Change Econ Dyn 52:221–243

Posner EA (2003) Economic analysis of contract law after three decades: success or failure? Yale Law J 112: 829–880

Schmidt KM (1998) Contract renegotiation and option contracts. In: Newman P (ed) The new Palgrave dictionary of economics and the law. Palgrave Macmillan, London

Schmitz PW (2001) The hold-up problem and incomplete contracts: a survey of recent topics in contract theory. Bull Econ Res 53(1):1–17

Schwartz A (1998) Incomplete contracts. In: Newman P (ed) The new Palgrave dictionary of economics and the law. Macmillan/Stockton Press, London/New York

Williamson OE (1985) The economic institutions of capitalism. The Free Press, New York

I

Independent Ethics Committee

► Institutional Review Board

Independent Judiciary

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

Abstract

After describing the closely related concepts of judicial independence and independent judicial review of policy, this entry offers an overview of four issues: (1) Reasons for establishing an independent judiciary, including its ability to resolve problems of information asymmetry between citizens – principals and public officials – agents, transform constitutional declarations to credible commitments and provide a mechanism of political insurance; (2) mechanisms for appointing judges and the jurisdiction of courts; (3) modeling the role of the judiciary as an additional veto player in games of collective decision-making and policy implementation; and (4) the judiciary as an explanatory variable and its effect on economic variables of interest like economic growth and the size of the government.

Definition

Judicial independence means that courts enforce the law and resolve disputes without regard to the power and preferences of the parties appearing before them (La Porta et al. 2004). Its theoretical antecedents are traced to the Enlightenment, and its application in practice dates to the US Constitution. Judicial independence is an indispensable part of the rule of law. The rule of law requires that laws apply equally to both ordinary citizens and public officials and that they protect the rights of individuals against the power of the state in both the political and economic spheres. In this respect the rule of law and judicial independence are inextricably linked with liberal democracy. The literature on the topic is enormous and cuts across different disciplines including law, economics, politics, and sociology. It is not possible to do justice to this scholarship in the confines of the present essay; rather its aim is to present a summary of the main issues. First, it considers the rationale of judicial independence and the closely related judicial review. Second, it looks at the institutional arrangements for judicial independence. Third, it considers how independent courts are modeled in the collective choice framework. Fourth, it discusses some evidence on the effects of judicial independence on economic variables of interest. These issues are analytically treated as separate but are best understood in relation to each other.

Rationale for Judicial Independence

Judicial Independence and Related Concepts

An independent judicial authority is necessary to resolve disputes and maintain the rule of law which are prerequisites for the functioning of a market economy and a free society. In general, two parties in dispute may resolve their differences by fighting violently against each other or by asking a third party to arbitrate. Realizing that fighting may result in serious inefficiency (destruction of life and property), they may ask a third party to adjudicate and agree to abide by its ruling. They will only do so, however, if they are reasonably confident that the adjudicating party is a neutral and unbiased referee. Disputes may emerge between private entities (citizens, companies, or other organizations), between private entities and the state which among other cases is always a litigant in cases of economic regulation and criminal acts, and between different state organizations (central government, local authorities, nationalized industries, and other public law bodies). Judges with the power to issue binding rulings must then be shielded from the threat of corruption and intimidation by both private litigants and the arms of the state.

However, the very act of referring a dispute to a mediator generates a new conflict: When the dispute resolver declares a winner and a loser, his legitimacy may be undermined leading to the collapse of the adjudication process and its benefits. The reason is that a ruling which obliges parties to behave in a particular way (take specific actions, pay damages, fines, sentence to prison) creates a two-against-one situation (winner and judge against the loser), which is resented by the loser. In order to overcome such problems, arbitrators base their rulings on generally accepted principles of justice and conduct as expressed in formal laws and informal norms and adopt rhetoric of normative justification. The two-against-one problem is even more pronounced in cases where the state is one of the disputants. A delicate balancing act then must be performed between the need to resolve disputes, protecting the independence of the judge, and ensuring that he is perceived as serving only notions of justice.

Closely related to judicial independence is the function of judicial review of policy, where courts may examine and subsequently ratify or annul laws and policy measures, passed by the legislature and enacted by the executive branch of government, for their compatibility with the constitution or other relevant statutes (like declarations of basic rights), and have been enacted according to the stipulated procedures (Stone Sweet 2002). Similarly to ruling in disputes between private parties, judicial review of policy is meaningless unless the reviewing judge is independent of the government. Two further related concepts are those of judicial activism and judicial discretion. Judicial activism is the propensity of courts to query the decisions of elected officials and range from "nullifying acts of the legislature, to abandoning neutral principles, to deciding cases in a politically 'liberal' or 'conservative' fashion" (Hanssen 2000, p. 538). Judicial discretion is understood as the degree to which the judiciary can implement rulings without being overruled by one of the other branches of government (Voigt 2008).

The interdisciplinary literature analyzing judicial independence can be divided into two strands. The first includes scholarship that treats a politically independent judiciary as an endogenous variable and examines the reasons for its establishment and the characteristics and degree of its political independence. The second strand considers courts as an explanatory variable and seeks to understand how it affects other political and economic variables of interest, mainly but not exclusively economic growth.

Reasons for Judicial Delegation

A number of in truth complementary explanations of an independent judiciary and its review powers have been proposed in the literature. For a review complementary to the issues taken up in the present section, the interested reader is referred to Harnay (2005). In all cases the starting point is the constitution. Constitutions, written or unwritten, and other fundamental charters specify the rules by which collective decisions are made and the constraints set upon the government and the citizens. They contain declarations of general principles, procedures, organizational forms, and rights and obligations, which, in the absence of complete information and perfect foresight about future changes in tastes and technology, are rendered as incomplete contracts riddled with problems of interpretation and enforcement. The judiciary is the arm that interprets and enforces the constitution, all ordinary laws, and policy measures, and for this reason judicial independence and judicial review are often analyzed jointly.

Modern scholarship uses the insights of the economic analysis of institutions and game theory to examine the benefits of delegation to courts; see Law (2009) and Tiede (2006). Delegation of decision-making by uninformed principals, like the citizens or their political representatives, to the judiciary, a specialized agent, offers several benefits. (a) It resolves problems of information asymmetry as courts develop the relevant expertise in resolving disputes and interpreting and enforcing the law, which subsequently allows specialization of labor and increases welfare. (b) By taking resolution of constitutional disputes away from partisan politics and handing it to "politically disinterested" judges, judicial independence promotes the long-run interests of citizens.

Politicians are better informed than ordinary citizens and exercise discretionary actions which opens up the opportunity for abuse of power to pursue their own interests at the expense of the rest of the society. Citizens may be protected from such abuses by subjecting politicians to elections, which allows voters to confirm or reject politicians, and by setting up checks and balances, where decision-making is divided between different arms of the government and each arm can block the actions of the rest. An independent judiciary is part of the latter mechanism. As it does not need to pander to short-term shifts in public opinion, it may be trusted to look after the long-run interests of citizens and control politicians. Thus, judicial independence is a mechanism that transforms constitutional declarations to credible commitments. Specifically, citizens wish to protect certain individual rights when the cost suffered by someone who is denied that right is very large relative to the gain obtained by others when the right is denied and when those who grant the right are uncertain whether they will be protected or harmed by that right (Mueller 1991). For example, pronouncements of individual freedoms, property rights, protection of minorities, nondiscriminatory taxation, and the like can be trusted by the citizens, who safe in this knowledge will develop longer time horizons increasing investment, growth, and welfare (Maskin and Tirole 2004). Note that in the credibility rationale of judicial delegation, the independent judiciary protects the interests of citizens against a mighty government and political competition and judicial review are substitutes. This is based on an underlying conflict between on the one hand politicians (who irrespective of partisan ideologies pursue their personal interests against those of the citizens) and on the other hand all the citizens.

(c) Contrary to the credibility view, the political insurance view of an independent judiciary considers the conflict between political groups competing for office and focuses on independent courts as a mechanism of political insurance. This approach builds on the famous thesis of Landes and Posner (1975) that a judiciary independent of the current legislature adds permanence to the distributive gains secured by the original winning political coalition. In essence the argument runs as follows (Stephenson 2003; Hanssen 2004a, b; Tridimas 2004, 2010). Constitutional judicial review implies that courts may prevent the election winner to implement his favored policy measures if found to violate the constitutional arrangements and the rights of citizens. However, in exchange for this constraint, when the same party is out of office, its opponent may also be prevented from implementing his favored policy. Hence, the losers of the political contest can use the review process as a mechanism to minimize the losses inflicted to them from the measures taken by the electoral winner. When political groups anticipate that they will not win every election and therefore they will be out of power, constitutional judicial review is a useful mechanism to restrain those in control of the government. Constitutional review by an independent judiciary lowers the risks associated with the uncertain outcomes of collective choice. On the above reasoning, one expects judicial review to be more pronounced in politics where political competition is strong and parties alternate in office. In the political insurance framework given the probability to win an election and the differences in the preferences of competing political groups, each political group is better off when an agent decides policy but prefers to delegate policy making to an "ally," that is, an agent which has preferences similar to its own. See Cooter and Ginsburg (1996) and Hayo and Voigt (2007) for an empirical investigation of the determinants of judicial independence.

Note that delegating thorny issues to the judiciary may offer short-run benefits to politicians who this way shift blame for unpopular decisions to independent agents to escape electoral punishment (Fiorina 1986). However, such delegation to the courts decreases the ability to claim credit for policies with a favorable impact. When the expected gains from shifting the blame exceed the expected losses from foregoing credit, the politician will choose to refer policy making to the judiciary. Shifting of responsibility is easier if the judiciary is perceived by the electorate as independent of the other branches of government. However, the latter view loses its explanatory

power when voters cannot be fooled and recognize the politicians' play.

Institutional Arrangements for Judicial Independence

Judicial review of the acts of government is the most politicized aspect of the behavior of courts. Judicial involvement in the political process and collective choice raises a fundamental question: Decision-making by an independent but unelected judiciary may go against deep-seated notions of majority decision-making and electoral accountability. As soon as discretionary powers are granted to the judiciary, a new principal–agent problem arises: A judiciary which is strong enough to block the legislative majority is also strong enough to pronounce rulings to pursue its preferences. What guarantees are there that the independent judiciary will not pursue its own interests at the expense of the citizens that it is supposed to protect? This is the well-known problem of "who will guard the guards?" going back to the ancient Greek philosopher Plato and the Roman poet Juvenal. Note the reverse dilemma too: accountability of the judiciary to give reasons and explain their actions is hardly controversial. But accountability by holding judges responsible for their decisions may infringe their independence, for it cannot be precluded that measures which aim to strengthen the accountability of an agent may be abused and weaken its independence. The solution of this dilemma depends on the arrangements that in practice balance the demands for judicial independence and accountability of the judiciary; see also Cappelletti (1983) and Shapiro (2002) for the tension between democracy and judicial independence. We divide the arrangements for judicial independence under two broad categories, structure and jurisdiction.

Structure

The political independence of judges increases with the following: (a) The smaller the involvement of the government in the process of their appointment and the larger the legislative majorities needed for their confirmation; independence is even higher when judges are nominated by the judiciary itself. (b) The longer their term of service; independence is also higher when judges serve for a single term only or do not seek reappointment. (c) The greater their financial autonomy which implies that salaries and budgets cannot be reduced by discretionary acts of the executive. (d) Transparency, the obligation to explain and justify rulings, enhances the independence of judges as it increases publicly available information, influences future courses of action, obliges nonelected judges to fully justify their decisions, and discourages politicians or other interested parties to intervene in judicial outcomes. However, there is no consensus regarding the optimal degree of transparency. For example, although full disclosure has a certain intuitive appeal, keeping secret the voting record of judges may also have some advantages in the specific circumstances of supranational judicial bodies like the European Court of the EU. The court does not disclose how individual judges have voted and, contrary to the US Supreme Court, dissenting opinions are not published. This secrecy protects judges against possible retribution from governments which lost their cases at the court. (e) The more difficult it is to discipline and dismiss judges. (f) The more difficult it is for the government to overturn judicial rulings it does not like. Rulings can be overturned by introducing new legislation or changing the status and power of courts. The latter is less likely when courts are bound to follow legal precedent and when their independence is declared in the constitution which, contrary to ordinary legislation, requires supermajorities to revise.

Jurisdiction

A range of issues are examined here – see Ginsburg (2002) for a detailed discussion of the structure and organization of the judiciary. (a) Whether ordinary courts can exercise judicial review of laws, as in the decentralized US system, or only specialized constitutional courts have such rights, as in the centralized system of the European countries following the model of the Austrian legal theorist H. Kelsen. (b) Whether judicial review is concrete or abstract. Under concrete the constitutionality of a law is checked in a case which is

actually litigated in front of a court, while under abstract a law may be examined without litigation. Related to this issue is whether review is carried out before or after the promulgation of a law. US courts practice concrete ex post review, while the French Constitutional Court offers an example of abstract a priori review. Abstract and a priori review is not based on a real case but on a hypothetical conflict and is conducted with less information about "facts" and as such is more limited but has the advantage that it can eliminate unconstitutional legislation before it actually does any harm. (c) In general, the power of the judiciary as an independent arm rises when individuals are granted more open access to the courts and ceteris paribus when courts are allowed to review more legislation, since under these circumstances the hold of the executive on policy making is weaker. Note however that easy access may encourage trivial applications for annulments frustrating the exercise of the will of the majority but also increasing the judicial workload and therefore the cost of the system. (d) The ability of court rulings to "make law." Although judiciaries do not make laws in the sense that legislatures do, insofar as they interpret legislation, their rulings become a source of law and bind future rulings, their independence is greater than otherwise. Similarly, by annulling those acts and measures that they find incompatible with the constitution and fundamental charters, courts have a form of negative lawmaking power.

Modeling the Behavior of an Independent Judiciary

The behavior of the judiciary in collective decision-making is studied by applying spatial decision models and game theory to the process of policy making. The judiciary is modeled as a rational agent pursuing an objective function defined over one or more policy variables and is pitted against the executive arm and the legislature in a sequential game; see Ferejohn and Weingast (1992), Hanssen (2000), Vanberg (2001), Rogers (2001), Tsebelis (2002), and Stephenson (2004). Introducing the judiciary in the collective choice game typically adds the highest court as a veto player, which affects the set of feasible alternatives against the status quo. In this policy game the executive and legislative branches are not only interested in the outcomes of their own decisions but also on whether their decisions will trigger the judiciary to move and reverse such decisions.

The literature makes two key assumptions about the preferences of the judges. First, their preferences are based on "deeply internalized" notions of justice, the rule of law, and respect for legal reasoning. Second, judges would like to see their rulings implemented. However, in modeling the utility function of judges, normative objectives are not specified. To a large extent this comes from the generality of the rule of law that eludes more specific normative specification. Application of the rule of law does not necessarily imply that a "good" law is applied. The law may privilege the interests of whomever the lawmakers wish to favor. As Shapiro (2002) put it, "The rule of law requires that the state's preferences be achieved by general rules rather than by discretionary-arbitrary-treatment of individuals" (p. 166). For example, courts of the apartheid era in South Africa were upholding the law of the land, but to the black population, they could hardly appear as neutral and independent arbiters between the interests of different races.

Empirical Studies of the Economic Effects of Judicial Independence

Application of the rule of law promotes a just society, protects individual rights, and defends citizens against predatory governments, advancing in turn economic goals. Major research breakthroughs were made with the construction of indicators of the independence and the power of the judiciary as represented by supreme courts. Empirical research has found that countries with higher degrees of judicial independence enjoy higher economic performance (Henisz 2000), greater economic and political freedom (La Porta et al. 2004), and a lower share of taxes (Tridimas 2005). Most interestingly, Feld and Voigt (2003) distinguish between de jure independence, as

described in legal texts setting up the supreme court of a country, and de facto independence which is independence of the supreme court of a country as it is actually implemented in practice, and find that only de facto judicial independence is conducive to growth.

Regarding the effect of the method of selecting judges, appointment or election, on judicial outcomes, the literature notes a selection effect (the ideological preferences of judges who are elected may differ from the preferences of judges who are appointed) and an incentive effect (judges seeking reelection are more sensitive to the preferences of the electorate). It is found that appointed judges are more independent than elected ones, since elected judges are more sensitive to electoral considerations and may attach greater weight to the interests of litigants from groups who are presumed to have large electoral power; see Hanssen (1999).

Informative as these findings may be, several open questions remain. In the first instance, there is the perennial problem of reverse causality, that is, richer countries can afford good judicial institutions rather than good judicial institutions leading to higher income. Second, the judiciary is approximated by the highest constitutional court and the latter is treated as a single decision taker. This ignores that in reality the judiciary comprises a hierarchy of lower and higher courts. In addition, even though the constitutional court itself is a collective body comprising several justices, subject to the well-known problems of reaching a collective decision, these problems are assumed away and the court is treated as a single decision taker. Finally, from the viewpoint of policy advice, the creation of legal institutions conducive to economic success requires long gestation periods, which may be of little comfort to a government facing pressing short-run demands for growth-promoting policies. Significantly, the results from reforming the courts of developing economies to be politically independent and introducing statutes incorporating principles and procedures of codes found in advanced western countries have been underwhelming (Carothers 2006). It appears that the same deep-lying factors of institutional failure were left intact and prevented the legal reforms to function as intended.

Cross-References

- ▶ Constitutional Political Economy
- ▶ Credibility
- ▶ Good Faith and Game Theory
- ▶ Incomplete Contracts
- ▶ Institutional Economics
- ▶ Political Economy

References

Cappelletti M (1983) Who watches the watchmen? A comparative study on judicial responsibility. Am J Comp Law 31:1–62

Carothers T (2006) Promoting the rule of law abroad: in search of knowledge. Carnegie, Washington, DC

Cooter RD, Ginsburg T (1996) Comparative judicial discretion: an empirical test of economic models. Int Rev Law Econ 16:295–313

Feld PL, Voigt S (2003) Economic growth and judicial independence: cross country evidence using a new set of indicators. Eur J Polit Econ 19:497–527

Ferejohn JA, Weingast BR (1992) A positive theory of statutory interpretation. Int Rev Law Econ 12:263–279

Fiorina M (1986) Legislator uncertainty, legislative control and the delegation of legislative power. J Law Econ 2:33–51

Ginsburg T (2002) Economic analysis and the design of constitutional courts. Theor Inq Law 3:49–85

Hanssen FA (1999) The effect of judicial institutions on uncertainty and the rate litigation: the election versus appointment of State judges. J Leg Stud 28:205–232

Hanssen FA (2000) Independent courts and administrative agencies: an empirical analysis of the States. J Law Econ Org 16:534–571

Hanssen FA (2004a) Is there a politically optimal level of judicial independence? Am Econ Rev 94:712–799

Hanssen FA (2004b) Learning about judicial independence: institutional change in the State courts. J Leg Stud 33:431–474

Harnay S (2005) Judicial independence. In: Backhaus J (ed) The Elgar companion to law and economics, 2nd edn. Elgar, Cheltenham, pp 407–423

Hayo B, Voigt S (2007) Explaining de facto judicial independence. Int Rev Law Econ 27:269–290

Henisz W (2000) The institutional environment for economic growth. Econ Polit 12:1–31

La Porta R, Lopez-de-Silanes F, Pop-Eleches C, Shleifer A (2004) Judicial checks and balances. J Polit Econ 112:445–470

Landes W, Posner R (1975) The independent judiciary in an interest-group perspective. J Law Econ 18:875–911

Law DS (2009) A theory of judicial power and judicial review. Georget Law J 97:723–801

Maskin E, Tirole J (2004) The politician and the judge: accountability in government. Am Econ Rev 94: 1034–1054

I

Mueller DC (1991) Constitutional rights. J Law Econ Organ 7:313–333
Rogers JR (2001) Information and judicial review: a signalling game of the legislative-judicial interaction. Am J Polit Sci 45:84–99
Shapiro M (2002) The success of judicial review and democracy. In: Shapiro M, Stone Sweet A (eds) On law, politics and judicialization. Oxford University Press, Oxford, pp 149–183
Stephenson MC (2003) When the devil turns. . .: the political foundations of independent judicial review. J Leg Stud 32:59–90
Stephenson MC (2004) Court of public opinion: government accountability and judicial independence. J Law Econ Organ 20:379–399
Stone Sweet A (2002) Constitutional courts and parliamentary democracy. West Eur Polit 25:77–100
Tiede LB (2006) Judicial independence: often cited, rarely understood. J Contemp Leg Issues 15:129–261
Tridimas G (2004) A political economy perspective of judicial review in the European Union. Judicial appointments rule, accessibility and jurisdiction of the European Court of Justice. Eur J Law Econ 18:99–116
Tridimas G (2005) Judges and Taxes: judicial review, judicial independence and the size of government. Const Polit Econ 16:5–30
Tridimas G (2010) Constitutional judicial review and political insurance. Eur J Law Econ 29:81–101
Tsebelis G (2002) Veto players: how political institutions work. Princeton University Press, Princeton
Vanberg G (2001) Legislative-judicial relations: a game theoretic approach to constitutional review. Am J Polit Sci 48:346–361
Voigt S (2008) The economic effects of judicial accountability: cross-country evidence. Eur J Law Econ 25:95–123

Independent Regulatory Authorities

Régis Lanneau
Law School, CRDP, Université de Paris Nanterre, Nanterre, France

Abstract

Independent regulatory agencies are now considered to be the sign of modern economic regulatory systems. They proliferated since the 1980s, and it is believed that they are enhancing the efficiency of regulation. In this entry, I will use law and economics to develop some rationale to explain their diffusion and emergence.

Introduction

Independent regulatory agencies represent one of the key features of modern economic regulation. They have indeed proliferated by a factor three to six between the end of the 1980s and the beginning of the 2000s in OECD countries (Jacobzone 2005; see also Pollitt and Bouckaert (2000), considering that the spread of the new public management doctrine is a key to understand the rise of agencies). Almost all OECD countries now have at least financial regulator, energy regulator, environmental agency, and telecommunication authority. Developing countries are following a similar trend since the beginning of the 1990s. In Latin America for example (but the same trend is observed in Asia), less than 45 authorities existed in 1979 (and mostly regarding financial regulation); in 2002, this number was multiplied by three to reach 138 (Jordana and Levi-Faur 2005). Independent regulators are now established in major infrastructure and economic sectors as well as social and environmental arena.

These (administrative) agencies, established in general by legislative acts, are entrusted with substantial (but variable) regulatory power – from rulemaking to adjudication and sanctions – and granted a certain level of independence (especially regarding the executive branch). This independence materializes, quite often, with fixed terms, limits regarding reappointment, guarantees against removal (a "cause" is required), a staffing structure allowing a significant place to experts (and not politicians), and collegiality at its head. This independence is never absolute since, after all, agencies are agents, acting on behalf of a principal (but since agents, they could of course pursue their own agenda, Moe 1990). Independence cannot then be the absence of accountability (see also Çetin et al. 2016; Maggetti 2012).

This "rise of the non-elected" (Vibert 2007) – and more generally of the regulatory state (Glaeser and Shleifer 2003) – is striking. It would certainly be possible to notice the influence of the European Union regarding the liberalization of certain economic sectors (an independent energy regulator was for example required) or the role of the world bank and its market-oriented

regulatory arrangement (see also Gilardi 2005). Nevertheless, the rationale (and legitimacy) for their emergence and diffusion is often consequentialist and law and economics could then be used to assess critically the validity of these rationale. If independent regulatory agencies exist, it is because they are supposed to lead to more net benefits than if the regulation was in the hands of politicians, judges, or mere dependent regulatory agencies.

In this entry, we will not provide a law and economics explanation for the rise as such of independent regulatory agencies (for development regarding this dimension, see, for example, Glaeser and Shleifer 2003; Law and Kim 2011). These explanations are of course not the only one and cannot explain, as such, the diversity of design, power, and functioning of independent regulatory agencies (see, for example, Eberlein 2000 on the absence in Germany of a sectoral regulator for electricity before the influence of the energy directives of the European Union).

Independent Regulatory Agencies as a Shield Against Interest Groups

Independent regulatory agencies are supposed to reduce the amount of political rent-seeking and its adverse effects regarding the design of policies (Shapiro 1988). Indeed, elected regulators, according to a basic economic logic, are eager to maximize their support function in order to get elected or reelected; they might not then have all the incentives to enact welfare enhancing regulations to "buy" the support of some special interest groups (Stigler 1971; Posner 1974; Peltzman 1976). Reducing or removing the political factors in regulation could then be seen as a relevant strategy. A variation around this theme can be found in Glaeser and Scleifer (2003): for these authors, regulatory agencies emerged to face the problem of large, deep-pocket firms that were able to manipulate the courts. Rent-seeking being considered in this variation at the level of courts.

This logic suffers from three problems:

First, even if the political factor is removed, the members of the regulatory agencies might have their own agenda which are not compatible with the "public interest" (i.e., obtaining a larger budget, extending their powers) and since they are "independent," it might be difficult to incentivize them to do the "right" thing, especially since their goals are often varied, mixed, broad, and unclear. The dilemma between independence and accountability is clear at this level. Moreover, from a strictly legal point of view, the possibility to delegate regulatory power delegate (and the extent of the delegation) to "independent" agencies is sometimes unclear (Veljanovski is considering them as "constitutional anomalies," Veljanovski 1991) and the necessity to coordinate the agency with other institution (e.g., for enforcement).

Second, the risk of capture and rent-seeking is not to be excluded by the mere fact of independence. Indeed, agencies' members are often experts in a field and as such have had the opportunity of repeated interaction with some of the firms they are regulating. Moreover, when they cannot be reappointed, they could wind-up in a high paying job in one of the companies they were regulating, hence incentivizing them to develop regulations not always inspired by the public interest.

Third, this logic also disregards the fact that independent agencies (and their design) could be the result of rent-seeking activities (because, for example, some industries could consider that it would be easier to pressurize this type of agency).

Take for example the Interstate Commerce Commission (ICC) spawned by the Interstate Commercial Act (ICA) of 1887. It is often considered as the first Independent regulatory agency. For Stigler (1971), railroads were able to limit of interstate trucking through their influence over the ICC. For Mullin (2000), the ICC was captured by the shippers, not the railroads. Indeed, railroads were refused to raise their rates despite the evidence of increasing input costs. Stock market evidence is pointing toward a capture by long-haul railroads at the expense of short-haul railroads (Prager 1989). Even if it is difficult to precisely identify which interest group "captured" the ICC, evidence is not showing that an "independent" regulatory agency could fully solve the problem.

Judicial review of regulatory agencies could reduce the problem, but not entirely solve it, since it is difficult to assess when the margin of discretion has been trespassed. If judicial review exists, the Glaeser and Shleifer understanding of the rise of regulatory agencies should be reinterpreted.

Independent Regulatory Agencies as a Tool to Reduce Decision-Making Costs

The idea behind this logic is simple: since the field which is regulated is perceived as a technical field (high information requirements), only experts are required to ensure an efficient regulation. Politicians or judges are, in some domains, ill-equipped to deal with the complexity and technicity required to design efficient public policies. Taking advantage of agency expertise could then make sense. This logic leads to distinguish between technical fields (choosing the best means to achieve a specific end, providing that there is a technical solution) and political – or nontechnical – fields (choosing the ends or the means when they are requiring more than a mere expertise). Such a distinction is not new but is crucial to understand the rise of independent regulatory agencies. A variation of this line of reasoning will stress the possibility to use these agencies for "blame shifting": if the domain is unpopular, delegating it to a regulatory agency could be an easy way to avoid the "blame" which could be the results of regulations (Epstein and O'Halloran 1999).

For this distinction to make sense, it is required to have the possibility to evaluate the output of these agencies. Indeed, since the problem is supposed to be technical, deviation should be easy to identify. Nevertheless, in the real world, this is far from being the case. For example, the Treaty on the Functioning of the European Union, in its article 127, states that: "The primary objective of the European System of Central Banks (hereinafter referred to as 'the ESCB') shall be to maintain price stability. Without prejudice to the objective of price stability, the ESCB shall support the general economic policies in the Union with a view to contributing to the achievement of the objectives of the Union as laid down in Article 3 of the Treaty on European Union." Regarding the first goal, the question of the possibility to use nonconventional tools like the Outright Monetary Transaction Program has been challenged (ECJ case C-62/14). Moreover, it is difficult to identify when a european central bank's action has been undertaken to support "the general economic policies in the EU." In other words, when a regulatory agency is assigned more than one mission, it is extremely difficult to assess its efficiency.

Independent Regulatory Agencies as a Way to Ensure Credibility of Long-Term Policy-Commitment

The third logic – temporal inconsistencies and credibility (Majone 1996) – was first pointed regarding monetary policy – and central banks – by Kyland and Prescott in 1977. As they specified, since "economic planning is not a game against nature but, rather, a game against rational economic agents" (Kydland and Prescott 1977), the problem is to ensure the credibility of announced rules (see also Elster 2000; Alesina and Tabellini 1988). If the government retains a discretionary power, it might not stick to a policy that was announced to further other agendas. Knowing this possibility, agent will react to the announced policy depending on its perceived credibility. In other terms, even if a policy has been designed to promote public interest, this policy might be rendered ineffective by rational actors who will anticipate some future move by policy makers. When this problem exists, one solution to ensure credibility would be to delegate the regulatory power to an independent regulatory agency (see also Majone 2001; Dixit 1996) whose purpose will only be to follow some established rules, reducing the probability of instrumental and political use. The rise of independent central banks is the best illustration of this "bootstrapping" logic. A variation of that logic is stressing the possibility to use regulatory agencies to avoid present politicians to bind future politicians and reduce uncertainties (Cukierman et al. 1992 for manipulation regarding the tax system).

Once again, the problem of the possibility to identify clearly the mandate of the institution to assess if it is trespassing its power remains. The possibility of a judicial review could help but not entirely solve the problem.

Conclusion: How to Insure Accountability of Independent Regulatory Agencies?

Ensuring the accountability of these nonelected bodies remains the central question in democratic societies. If it were possible to evaluate the performance of these agencies, accountability would be easy to assess. Nevertheless, this task is rarely undertaken (Gilardi and Maggetti 2010) and probably too difficult to be a viable solution. Of course, some papers are trying to assess the impact of these regulatory agencies (quite often only indirectly, for example, Jakee and Allen 1998) but it appears difficult to disentangle regulation (and their perceived efficiency) from regulatory agencies (and their specific role as an institution of regulation). Moreover, even if some players seemed to have benefited from a regulation enacted by regulatory agencies, it remains difficult to be fully certain that it leads to negative effects (considering the difficulty to identify what would be the best achievable). As Ronald Coase advocated, it is required to compare different institutional framework not in the abstraction of a model but with an eye on what is achievable (Coase 1960).

Independence and accountability could only be ensured through a system of a new separation of power which design remains to be identified (and game theory is thus required, Hägg 1997). Among the leads and good practices, a strict adherence to the rule of due process and a perfect transparency are certainly required, so is a system of appeals of their decisions (see the entry on ▶ Regulatory Impact Assessment). Moreover, a special attention should be paid to agencies coordination (at the level of the European Union, a system of energy regulator coordination was mandated considering the specific dimension of the electricity and gas market. It led to the creation of the European Regulators' Group for Electricity and Gas. Moreover, a dialogue between specialized regulatory agencies and competition authorities is also crucial for efficient regulation). Nevertheless, designing a perfect system is a Herculean task requiring information on the full constraint system (both legal and social); the best we can do is only to reduce the probability of blatantly inefficient regulations.

Cross-References

- ▶ Public Choice: The Virginia School
- ▶ Regulatory Impact Assessment
- ▶ Rent Seeking
- ▶ Separation of Power

References

Alesina A, Tabellini G (1988) Credibility and politics. Eur Econ Rev 32:542–550

Çetin T, Sobaci MZ, Nargeleçekenler M (2016) Independence and accountability of independent regulatory agencies: the case of Turkey. Eur J Law Econ 41(3):601–620

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

Cukierman A, Edwards S, Tabellini G (1992) Seignorage and political instability. Am Econ Rev 82(3):537–555

Dixit A (1996) The making of economic policy. A transaction-cost politics perspective. The MIT Press, Cambridge, MA

Eberlein B (2000) Institutional change and continuity in German infrastructure management: the case of electricity reform. Ger Polit 9(3):81–104

Elster J (2000) Ulysses and the Sirens. Studies in the rationality and irrationality. Cambridge University Press, Cambridge

Epstein D, O'Halloran S (1999) Delegating powers: a transaction cost politics approach to policy making under separation of powers. Cambridge University Press, Cambridge

Gilardi F (2005) The institutional foundations of regulatory capitalism: the diffusion of independent regulatory agencies in Western Europe. Ann Am Acad Pol Soc Sci 598:84–101

Gilardi F, Maggetti M (2010) The independence of regulatory authorities. In: Levi-Faur D (ed) Handbook of regulation. Edward Elgar, Cheltenham, pp 201–214

Glaeser EL, Shleifer A (2003) The rise of the regulatory state. J Econ Lit 41(2):401–442

Hägg PG (1997) Theories on the economics of regulation: a survey of the literature from a European perspective. Eur J Law Econ 4(4):337–370

Jacobzone S (2005) Independent regulatory authorities in OECD countries: an overview. In: OECD (2005) Designing independent and accountable regulatory authorities for high quality regulation. https://www.oecd.org/gov/regulatory-policy/35028836.pdf
Jakee K, Allen L (1998) Destructive competition or competition destroyed? Regulatory theory and the history of Irish road transportation legislation. Eur J Law Econ 5(1):13–50
Jordana J, Levi-Faur D (2005) The diffusion of regulatory capitalism in Latin America: sectoral and national channels in the making of a new order. Ann Am Acad Pol Soc Sci 598:102–124
Kydland F, Prescott EC (1977) Rules rather than discretion: the inconsistency of optimal plans. J Polit Econ 85(1):73–491
Law M, Kim S (2011) The rise of the American regulatory state: a view from the progressive era. In: Levi-Faur D (ed) Handbook of regulation. Edward Elgar, Cheltenham, pp 113–128
Maggetti M (2012) The media accountability of independent regulatory agencies. Eur Polit Sci Rev 4(3):385–408
Majone G (1996) Temporal consistency and policy credibility: why democracies need non-majoritarian institutions, EUI working paper, RSC no. 96/57, European University Institute, San Domenico di Fiesole
Majone G (2001) Two logics of delegation. Agency and fiduciary relations in EU governance. Eur Union Polit 2(1):103–121
Moe TM (1990) The politics of structural choice: towards a theory of public bureaucracy. In: Williamson OE (ed) Organization theory: from Chester Barnard to the present and beyond. Oxford University Press, New York, pp 116–153
Mullin WP (2000) Railroad revisionists revisited: stock market evidence from the progressive era. J Regul Econ 17(1):25–47
Peltzman S (1976) Toward a more general theory of regulation. J Law Econ 19:211–240
Pollitt C, Bouckaert G (2000) Public sector management reform: a comparative analysis. Oxford University Press, Oxford
Posner RA (1974) Theories of economic regulation. Bell J Econ Manag Sci 5:335–358
Prager RA (1989) Using stock price data to measure the effects of railroad regulation: the Interstate Commerce Act and the railroad industry. RAND J Econ 20(2):280–290
Shapiro M (1988) Who guards the guardians? Judicial control of administration. University of Georgia Press, Athens
Stigler GJ (1971) The theory of economic regulation. Bell J Econ Manag Sci 2:3–21
Veljanovski C (1991) The regulation game. In: Veljanovski C (ed) Regulation and the market. IEA, London
Vibert F (2007) The rise of the non-elected, democracy and the new separation of power. Cambridge University Press, Cambridge

Inference

▶ Rationality

Informal Economy

▶ Shadow Economy

Informal Law

▶ Customary Law

Informal Sector

Ozan Hatipoglu
Department of Economics, Bogazici University, Istanbul, Turkey

Abstract

I review and discuss definitions of informal sector introduced by social scientists over the last half century. I describe how informal institutions, informal markets, and their participants' activities form together the informal sector. I provide insight into the difficulties in defining informal sector from a judicial point of view. I discuss causes and consequences as well as economic costs and benefits of the informal sector. Finally, I provide a brief analysis of the existing econometrics methods in measuring its size.

Synonyms

Hidden Economy; Irregular Sector; Parallel Economy; Shadow Economy; Underground Economy

Definition

All income generating activities outside the regulatory framework of the state.

Definitions

A Brief History of Definitions

A formal meaning to informal sector (A plethora of adjectives are used to describe the informal sector in common language, e.g., shadow underground, subterranean, hidden, parallel, or irregular.) was not given until International Labor Organization (ILO) first officially introduced the term in its study of urban labor markets in developing countries (Hart 1973; Chaudhuri 2010). It was categorized as part of the urban labor force, which operates outside of formal labor regulations such as wage contracts or social security laws. The main emphasis in this conceptualization was on self-employed urban workers. ILO extended the definition of the term in 2002 to include also unregistered or unprotected labor working in formal sector firms. While this newer definition emphasizes labor markets, where most of the informal activity takes place, it is far from being authoritative or unique in its scope. In fact, a consensus on the definition of the informal sector is difficult to find because its coverage is nested in several branches of social sciences due to the multidimensional activities involved (Labor Economics, Anthropology, Sociology, Psychology, Finance, Macroeconomics, Criminology, and Statistics provide different methodological approaches to the analysis of the informal sector.) (Gërxhani 2004).

Early works on informal sector have associated it with the existence of a dual labor market in developing countries (Hart 1973; De Soto 1989; Tokman 1972, 1978) and considered it as synonymous for small and self-employed. These were the key characteristics that distinguished it from formal sector where wage-employment and a well-regulated labor market existed together. This is hardly surprising given the dualistic nature of most of the world economies in the postwar and the postcolonialist era. A developed urban economy existed together with a subsistence economy based on agriculture. The rise of urban industries caused massive rural-urban migration that increased the available labor force in cities but at the same time failed to generate enough employment. The surplus labor was forced to generate its own means of survival, and thereby it created a secondary urban employment sector that became later also an entry point for fresh immigrants (Mazumdar 1976) (Although wages in the informal sector are less than those in the formal sector in most cases, there is a wide diversity of earnings within the informal sector as ILO's Regional Employment Program, known by its Spanish acronym PREALC, reports (Tokman 1990; Mazumdar 1976; Lemieux et al. 1994)).

The dualist approach employed by researchers in treating informal sector as a separate entity received major criticism in early 1990's because of a buildup in empirical findings documenting close linkages between formal and informal sectors. Portes and Schauffler (1993) point to the existence of microproducers capable of producing with modern technology and capital accumulation, but also argue that this type of informal production is an exception in many developing countries including Latin America. Erase Harris (1990) and Sethuraman and Maldonado (1992) rejects both reject the dualist approach on the grounds that linkages between these two sectors play a greater role in explaining their formation than migration. This new strand of literature also suggests that entry to the informal sector might be due to individual choices, which are influenced by barriers to access such as heavy regulation, high taxes, or social factors.

A distinction between the legal and illegal nature of the informal work wasn't made until last two decades when authors from a variety of disciplines (Castells and Portes 1989; De Soto 1989; Feige 1990; Harding and Jenkins 1989) started to define informal sector commonly as that *all income generating activities outside the regulatory framework of the state.* This definition has now become the most widely used characterization of the informal sector in economics. Chen (2004) argues that during the 1990s, the emergence of the concepts of social capital and social networks led to questioning of the value of the concept of informality "even by its main proponents – and a significant decline in its use" (Klein 1999; Hart 1995; Portes 1994).

Since the beginning of the twenty-first century, the discussion on the definition of informal sector

has moved away from epistemological concerns to more practical issues such as how to transform informal sector concepts into instruments for statistical measurement and public policy purposes. Currently, the ILO terminology explains informality around three main notions: the *informal sector* refers to production and employment in unregistered enterprises; *informal employment* focuses on employment outside of the labor protection regulations of a given society, both in formal and informal firms; and *informal economy* covers all firms, workers, and institutions that operate outside the legal regulatory framework of society and their revenue-creating activities.

Labor Market-Based Definitions

Dual labor market theories (Doeringer and Piore 1971; Saint-Paul 1997) divide the labor market into primary and secondary or informal and illegal sectors. The primary sector consists of regular, wage jobs that are regulated and taxed and come with well-defined contracts and social security benefits. Secondary sector provides jobs with lower wages and less regulation, such as those in the service sector. Jobs in this sector are also sometimes referred to as pink-collar jobs. Informal sector consists of mostly self-employed who cannot access primary or secondary sectors. They work by themselves or create small unofficial business units in which workers are hired and transactions are made off-the-books and payments are cash-only. The final category is the illegal sector and it consists of criminal activities that generate income.

H. De Soto (1989) emphasizes the regulatory framework where the main distinction between informal and formal sector is the legal status of the activities. The legal status of the informal sector activities is a gray area in common law and a distinction is almost always necessary between what is legal and what is not. For example, a production process might be illegal in the sense that it avoids or circumvents work regulations while the output resulting from this process might be not.

Most studies on informal sector exclude criminal activity such as drug-trafficking or human trade from a labor market-based definition of the informal sector. Unlike the criminal sector, the output in the informal sector is legal. Undeclared production of officially recorded firms is also generally considered as part of the informal sector. However, when the labor market-based definition is extended to include official firms, i.e., irregular sector, legal status of the informal sector is not straightforward. The production process in the irregular sector might be plagued with illegal procedures, such as tax evasion, avoidance of labor safety regulations, and social security fraud. Informal sector, therefore, can also be conceptualized as a market where illegal production of legal goods by unofficial firms takes place.

One can further refine this classification by distinguishing between activities that use monetary transactions versus those that use nonmonetary transactions. Illegal informal activities that use monetary transaction include trade with stolen goods, drug dealing. Another common illegal informal activity is money laundering which is defined as the set of actions taken by individuals or organizations to cloak the source of funds that are created by criminal activity (Masciandar, forthcoming). Illegal informal activities that use nonmonetary transactions comprise of barter of drugs and stolen goods, theft for own use, as well as producing drugs for own use. Regular informal activities that use monetary transactions can be categorized into *tax evasion* such as unreported income from self-employment, wages, salaries, and assets from informal work and into *tax avoidance*, such as employee discounts and other benefits.

Study-Based Definitions

Since there is a substantial body of literature on informal sector descriptions, it is customary for researchers to define informal sector in accordance with the particular research question at hand. For example, studies trying to assess the size of the informal sector provide definitions of informal sector that are directly or indirectly measurable even though they might include activities that are nontaxable. Smith (1994) puts one such working definition forward as "market based production of goods and services, whether legal or illegal that escapes detection in the official

estimates of Gross Domestic Product (GDP)." Another instrumental definition with a taxability condition is provided by Schneider (2006) as "all income from monetary or barter transactions of legal goods and services that would be taxed if it were reported to tax authorities."

Most statistical measurement studies exclude criminal activities because the magnitude of the transactions involved are very difficult to measure, but they are also broad enough to cover the diversity of ways in which the informal sector reveals itself in different countries. Studies concentrating on tax evasion or firm regulation focus on irregular sector rather than informal labor markets and take the informal sector to be a place of shadow activities of official firms.

In similar vein, the definition and the scope of the informal sector might also differ depending on whether the focus of the study is a developing or a developed country. In the former case, informal sector studies generally concentrate on labor-intensive informal work, whereas in the latter researchers have focused mostly on tax evasion and optimal regulation.

Informal Institutions, Informal Markets, Participants, and Activities

Informal institutions are governing arrangements that are created and sanctioned outside the regulatory reach of the state (Chen 2004). They are based on traditional power hierarchies or other socially accepted norms. Informal institutions as informal sector participants are more prevalent in developing countries when compared to developed countries.

Informal *markets* are organizational arenas in which factors of production and goods and services are traded.

Informal activity can refer to various informal dimensions of economic activity in an economic system. One way to describe this broad range of activities is to classify them with respect to market participants undertaking them. Individuals, small- or large-scale corporations, formal and informal institutions, financial and nonfinancial institutions, and governments all can engage in informal activities in a variety of ways. In developed countries, formal institutions play a larger role in informal sector activities when compared to informal institutions. A corporation, for instance, can engage in informal sector activity by:

(i) Evading taxes through hiding income from its legal activities
(ii) Hiring informal workers
(iii) Avoiding regulations with the aim of tax avoidance or reducing production costs to increase pretax profits
(iv) Making hidden transactions with smaller informal firms that provide cheaper inputs in the lower end of the supply-chain.
(v) Engaging in illegal activities such as bribery to evade taxes or to avoid workplace or environmental regulations

Individuals or households can also engage in informal activities. Individuals might work as self-employed without registration, or they might work for an informal employer such as own-account informal employers or informal own-family enterprises. Household nonmarket production is a type of informal activity that is not regulated by the state unless it is done for others. Since own household production does not create income it is generally left out from measurement studies and considered as a separate category outside the formal/informal sector divide.

Official state bodies such as governments, municipalities, or state departments can also engage in shadow activities in a variety of ways. They might extend work to informal employers without official tenders. State officials might accept bribery to facilitate procedures or to keep a blind eye on informal activities whether they are legal or not (Choi and Thum 2004; Dreher and Schneider 2006).

The formal financial institutions also engage in shadow activities in several ways. International Financial Stability Board (FSB) (2014) reports that shadow banking now accounts for a quarter of the global financial system. In many developing countries, the volume of informal finance lead by shadow banking exceeds formal finance in

terms of assets and volume of transactions. In developed countries, prior to the financial crisis of 2008, shadow banking mostly referred to legal structures that are used by official banks to keep complicated and risky securities off their balance sheets. With increased regulation on bank operations, informal banking now has become an independent sector and its definition is now extended as "credit intermediation involving entities and activities outside the regular banking system."

Where formal banks have access to ample funds but are unable to control the use of credit, informal lenders can prevent nondiligent behavior but often lack the needed capital.

Formal Versus Informal: Causes and Consequences

In economics literature, taxes, social security burden, increased regulation, forced reduction in weekly working time, earlier retirement, unemployment, decline of civic virtue, and loyalty towards public institutions are counted among the main reasons for the existence of the informal sector (Frey and Hannelore 1983). Psychological studies offer factors such as tax morale and perceived fairness of the tax system and micro-sociological studies offer exclusion and other social barriers to entry as main causes of the informal sector.

In empirical studies, researchers have looked at the correlates of informal sector size by controlling other factors. Common significant factors found in these studies reflect the generally accepted aforementioned reasons, but some puzzles still exist (Hatipoglu and Ozbek (2011) report that a high informal sector size is associated with little redistribution and lower taxes especially in developing countries).

Monitoring and enforcement in the irregular sector are costly efforts for the regulator due to the high number of qualified personnel hours required. Stronger enforcement and heavier regulations can result statistically in less number of firms evading taxes, and a higher tax collection rate, but at the same time it may lead incumbent and newcomer firms to engage more in informal activities thereby increasing informal sector size and reducing tax revenues.

The same argument applies to informal labor markets. When faced with higher taxes, individuals might opt out from formal sector and switch to informal sector to protect their after-tax income. As a result, the formal sector might shrink in size as well as in productivity as more skilled people switch to informal sector. Depending on its workforce's skill distribution, a country might have a large informal sector with plenty of skilled jobs or relatively small informal sector with relatively high overall unemployment. This trade-off is analyzed by Ihrig and Moe (2004) that shows how enforcement policies and tax rates interact to determine the size of the informal sector. They find even small changes in tax rates can significantly affect the size of informal employment. On the contrary, increased enforcement has negligible effect on informal employment. Their results suggest that modest reductions in tax rates combined with modest punishment for tax evasion are more effective then strong enforcement alone. The size of the informal sector and the level of redistribution is the result of the joint effects of distributional factors and how lucrative the informal sector is.

Many authors argue that reducing costs of entry to formal sector can reduce the size of the informal sector and improve labor market performance. Increasing enforcement may also reduce informality but can have negative effects on unemployment and welfare such that there is a trade-off between lower informal employment and higher unemployment rate. However, the trade-off disappears when one prefers policies that aim at reducing the costs of being formal, as opposed to policies that increase the costs of being informal.

Many scholars have emphasized the costs of operating in both sectors (Loayza 1996; De Soto 1989, Tokman 1990; Sarte 2000; Portes et al 1989). In the formal sector, in addition to having to pay direct or indirect taxes, formal firms need to comply with guidelines such as labor safety, environmental protection, consumer safety and quality control, as well as administrative procedures such as registration, and other

bureaucratic paper work. Wage regulations and project delays stemming from contract negotiations with organized unions also increase formal sector's operating costs. Whereas in the informal sector, there are costs to getting caught since the informal firm is penalized if detected. The major cost or lack of benefit associated with the informal sector is, however, that its participants cannot take advantage of state-provided public goods such as the police, the judicial system, as well as of business-related resources such as those provided by official trade organizations.

Working in the formal sector requires a minimum set of skills and entry costs. Guilds and business associations require their members to pay annual fees and comply with regulations. In return, individuals have access to resources and infrastructure that can help to improve their productivity. Depending on the tax burden, the intensity of regulations in the formal sector and wages in the informal sector, formally trained individuals such as doctors and lawyers might also choose to work in the informal sector.

Informal Sector: Costs Versus Benefits

The existence of the informal sector can be thought of as the expression of negative sentiments of individuals who are overburdened by the state and choose the exit option. If, in fact, the informal sector is caused by the burden of regulation, then an increase in informal sector size can cause a vicious cycle by eroding the tax base and thereby causing a further increase in tax rates (Olson 1982). A large informal sector size can create problems for both the voters and the policy-makers in taking the optimal decisions because official indicators on unemployment, income, and consumption become misleading.

Depending on how lucrative the informal sector is and how productive the formal workers are, people may choose to work in the informal sector to accommodate for their low levels of productivity. This in turns leads to less output in the formal sector with little to redistribute. This is consistent with the observation that the existence of a large informal sector coincides with less redistribution in many developing countries. Hatipoglu and Ozbek (2011) show that informal sector may act as a redistributive mechanism in countries where the democratic rule is skewed towards the rich. Increasing the wages in the informal sector vis-a-vis the earnings in the formal sector leads to a decrease in subsidies and an increase in the informal sector size, which suggests that nontaxation is an indirect form of subsidizing incomes and that informal sector acts as an alternative way of redistribution.

Informal sector can also operate in a way to reduce the burden of taxation further by creating new jobs for the low skilled who otherwise would be dependent on the government. Empirical evidence from ILO and World Bank studies show that most of new jobs in the developing world over the past 15–20 years have been created in the informal sector. The main reasons are rapid growth of the labor force, insufficient formal sector job creation, and lack of social safety nets (Chen 2004; Blunch et al. 2001: 10; ILO 1972). In addition, most of the income created in informal sector is immediately spent in the formal sector; thus increased informal activity can also help the formal sector to grow.

Measuring the Size of the Informal Sector

Measurement of the informal sector size is one of the central themes of the informal sector literature. Correct measurement of the informal sector size is important for a variety of reasons. Firstly, it can identify the exact amount of distortion caused by informal sector activities in major national statistics such as GDP allowing both voters and policymakers to make more informed decisions. Secondly, causes and consequences of the existence of informal activities can be easier to identify and policy prescriptions can be made more efficiently.

Currently, there are a number of alternative estimation methodologies that differ in their approach in tackling the unobservable nature of the informal sector (See Frey and Hannelore (1984), Schneider (2006) and Schneider and

Enste (2000) for a more detailed presentation and critique of the available estimation methods).

Direct Methods

Direct approaches collect data directly from participants either through *microsurveys* or use the available data collected through *tax audits*. Since direct approaches depend on the respondents' willingness to comply it suffers from a statistical measurement error called sample bias. People who chose to respond are less likely to engage in informal activities and if they do, they tend to underreport those activities. This makes estimating the range of informal activities as well as translating it into monetary terms difficult. On the other hand, this approach provides detailed information about the structure of the informal sector that is difficult to obtain using indirect approaches.

In the *survey method*, a sample population thought to represent the actual population of a country is chosen to be respondents. The answers are statistically extended to the whole population. The *tax audit* method, on the other hand, measures the discrepancy between incomes declared for taxation and incomes measured by selective checks. This approach provides more precise estimates of the informal activities that selected respondents undertake, but it also suffers from a sample bias problem, since respondents are not selected randomly but based on their previous tax record and likelihood to evade taxes. Finally, except hidden income, most informal sector activities are poorly captured by this method.

There are two common deficiencies of the aforementioned direct approaches. Firstly, they only provide point estimates in time and do not provide information on how informal sector evolves over time in terms of size. Secondly, they provide also poor estimates of the composition of informal sector.

Indirect Methods

Indirect or *indicator* approaches use aggregate macroeconomic data or other "indicators" to infer about the size of the informal sector. These indicators are thought to capture the traces left by the informal sector activities in national statistics. One such trace is the *discrepancy between the official and the actual labor force*. A decline in labor force participation is an indicator for increased activity in the informal sector. The main disadvantage of this method is that changes to labor participation rates might be due to other reasons than increased informal activity. Another indicator of the informal activity is the discrepancy *between GNP measures based on expenditure and income*. Since GNP based on expenditure and income approach should be equal, any discrepancy can signal the size of the informal sector. National statisticians are likely to minimize this discrepancy because they are caused by measurement error.

Currency demand approach assumes that informal sector transactions are mostly handled in cash, and therefore excess money holdings above what money demand regressions would predict can capture the size of the informal sector. It is one of the most commonly used methods in OECD countries. Since not all transactions are paid in cash, this method might underestimate the size of the informal sector. Another weakness of this method is that the rise in currency demand might be due to the slowdown in demand deposits instead of a rise in informal activity. This method also assumes the same velocity for money in both sectors, which might not be true (See Tanzi (1983) for an introduction and Thomas (1999) and Schneider (2002) for criticisms of the method. Giles (1999) and Bhattarchya (1999) address some of those criticisms.).

Physical input method assumes all kinds of production use a general-purpose, measurable input such as electricity. Since economic activity and electricity consumption are highly correlated in data, one can infer about the size of the informal sector by comparing the measured GDP and imputed GDP calculated using the measured inputs. The main criticism against this method is that some informal activities do not use electricity at all, e.g., personal services and many activities use energy sources other than electricity, e.g., oil and natural gas. Finally, the efficiency of electricity usage might change over time such that estimations over time become inconsistent.

Transactions method uses Fisher's quantity equation that relates the volume of transactions

to the velocity of money, prices, and money supply Feige (1986). Official GDP is subtracted from the total volume of transactions implied by the equation. Here, assumptions have to be made about the velocity of money being constant in both sectors as well as the existence of a base year with no informal sector, both of which might fail to hold.

Econometric Modeling Methods

This approach treats the informal sector as an unobservable variable and uses structural equation modeling to control for the causal relations between the informal sector size and its causes (burden of taxation, regulation, etc.) as well as the informal sector size and its effects (money demand, labor force participation rate, etc.). Schneider (2004), for example, uses dynamic multiple indicators multiple causes (DYMIMIC) method to estimate the size of the shadow economies around the world. This method provides only relative estimates; therefore results have to be combined with absolute measures from currency approach to compute absolute measures for all countries.

Informal Sector Sizes in the World

There are many studies which estimate the size of the informal sector for large sets of countries using aforementioned methods. A detailed survey is beyond the scope of this article. Interested readers can check Zilberfarb (1986), Lacko (1996), Tanzi (1999), Thomas (1999), Schneider and Enste (2000), Chatterjee et al. (2006), and Schneider (2009) among others in chronological order.

Conclusion and Future Directions

Works on the informal sector have concentrated on its definition, measurement, causes and consequences, as well as its links to other spheres in law and economics. Latest economic realities such as free trade and market friendly reforms seem to have contributed significantly to the growing prominence of the informal sector in economic activities. The developmental history of third world countries as well as boom-bust cycles of the developed world have radically changed perceptions about the relationship between the formal and informal sectors and their roles in economic development. Researchers from a wide variety of backgrounds continue to explore links between the informal sector and other fields such as law, trade, finance, political economy, growth, and income distribution.

Within this realm, the role of informal finance has significantly grown during the last half-decade and especially after the financial crisis of 2008. One can expect to see more studies on the role of regulations in the financial industry in shadow banking as well as in other types of informal finance. Further contributions aside mainstream economics can be expected from network analysis, new institutional economics, microfinance, legal pluralism studies, as well as Foucauldian studies on political economy.

References

Bhattacharrya DK (1999) On the economic rationale of estimating the hidden economy. Econ J 109:348–359

Blunch N-H, Canagarajah S, Raju D (2001) The informal sector revisited: a synthesis across space and time. Social Protection Discussion Paper 0119. World Bank, Washington, DC

Castells M, Portes A (1989) World underneath: the origins, dynamics, and effects of the informal economy. The informal economy: studies in advanced and less developed countries, 12

Chatterjee S, Chaudhuri K, Schneider F (2006) The size and development of the Indian shadow economy and a comparison with other 18 Asian countries: an empirical investigation. J Dev Econ 80(2):428–443

Chen M (2004) Rethinking the informal economy: linkages with the formal economy and the formal regulatory environment. Paper presented at the EGDI-WIDR Conference unleashing human potential: linking the informal and formal sectors, Helsinki

Choi J, Thum M (2004) Corruption and the shadow economy. Int Econ Rev 12(4):308–342

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

Dreher A, Schneider F (2006) Corruption and shadow economy: an empirical analysis. Discussion Paper, Department of Economies, University of Linz

Feige EL (1986) A re-examination of the "Underground Economy" in the United States. IMF Staff Pap 33(4):768–781

Feige EL (1990) Defining and estimating underground and informal economies: the new institutional economics approach. World Dev 18(7):989–1002
Frey BS, Hannelore W (1983) Bureaucracy and the shadow economy: a macro-approach. In: Hanusch H (ed) Anatomy of government deficiencies. Springer, Berlin, pp 89–109
Frey BS, Hannelore W (1984) The hidden economy as an "unobserved" variable. Eur Econ Rev 26(1):33–53
Gërxhani K (2004) The informal sector in developed and less developed countries: a literature survey. Public Choice 120(3–4):267–300, 09
Giles D (1999) Measuring the hidden economy: implications for econometric modelling. Econ J 109(456): 370–380
Harding P, Jenkins R (1989) The myth of the hidden economy: towards a new understanding of informal economic activity. Open University Press, Milton Keynes
Hart K (1973) Informal income opportunities and urban employment in Ghana. J Mod Afr Stud 11(1):61–89
Hart K (1995) L'entreprise africaine et l'economie informelle. Reflexions autobiographiques. In: Ellis S, Faure Y-A (eds) Entreprises et entrepreneurs africains. Karthala/ORSTOM, Paris
Hatipoglu O, Ozbek G (2011) On the political economy of the informal sector and income redistribution. Eur J Law Econ 32(1):69–87
Ihrıg J, Moe K (2004) Lurking in the shadows: the informal sector and government policy. J Dev Econ 73:541–557
International Labor Organization (ILO) (1972) Employment, income and equality: a strategy for increasing productivity in Kenya. International Labor Organization (ILO), Geneva
Klein A (1999) The barracuda's tale: trawlers, the informal sector and a state of classificatory disorder off the Nigerian coast. Africa 69.04:555–574
Lacko M (1996) Hidden economy in east European countries in international comparison. International Institute for applied systems analysis working paper
Lemieux T, Fortin B, Frechette P (1994) The effect of taxes on labor supply in the underground economy. Am Econ Rev 84:231–254
Loayza NV (1996) The economics of the informal sector: a simple model and some empirical evidence from Latin America. Carn-Roch Conf Ser Public Policy 45: 129–162
Maldonado C, Sethuraman SV (1992) Technological capability in the Informal sector: metal manufacturing in developing countries, vol 22. International Labour Organization
Mazumdar D (1976) The urban informal sector, world development. Elsevier 4(8):655–679
Olson M (1982) The rise and decline of nations. Yale University Press, New Haven – Business & Economics – 273 p
Portes A (1994) The informal economy and its paradoxes. In: Smelser NJ, Swedberg R (eds) The handbook of economic sociology. Princeton University Press, Princeton, pp 426–447
Portes A, Castells M, Benton LA (eds) (1989) The informal economy: studies in advanced and less developed Countries. Johns Hopkins University Press, Baltimore
Portes A, Schauffler R (1993) Competing perspectives on the Latin American informal sector. Popul Dev Rev 19:33–60
Saint-Paul G (1997) Economic integration, factor mobility, and wage convergence. Int Tax Public Finance Springer 4(3):291–306
Sarte PD (2000) Informality and rent-seeking bureaucracies in a model long-run growth. J Monet Econ 46:173–197
Schneider F (2002) Size and measurement of the informal economy in 110 countries around the world. Working paper. Department of Economics, Johannes Kepler University of Linz
Schneider F (2004) The size of the shadow economies of 145 countries all over the world: first results over the period 1999 to 2003. EZA discussion paper 1431
Schneider F (2005) Shadow economies around the world: what do we really know? Eur J Polit Econ 21(3): 598–642
Schneider F (2006) Shadow economies of 145 countries all over the world: what do we really know? Johannes Kepler University of Linz, Manuscript
Schneider F (2009) Size and development of the shadow economy in Germany, Austria and Other OECD-countries. Some preliminary findings. Revue économique Presses de Sciences-Po 60(5):1079–1116
Schneider F, Enste D (2000) Shadow economies: size, causes, and consequences. J Econ Lit 38(1):77–114
Soto H (1989) The other path: the invisible revolution in the Third World. Harper Row, New York
Smith P (1994) Assessing the size of the underground economy: the Canadian statistical perspectives. Canadian Economic Observer 3(11-010):16–33
Tanzi V (1983) The underground economy in the United States: annual estimates, 1930–80. Staff Papers, International Monetary Fund 30:283–305
Tanzi V (1999) Uses and abuses of estimates of the underground economy. Econ J 109(456):338–340
Thomas JJ (1999) Quantifying the black economy: 'Measurement without theory' yet again? Econ J 109(456): 381–389
Tokman V (ed) (1972) Beyond regulation: the informal economy in Latin America. Lynne Rienner Publishers, Boulder
Tokman V (1978) An exploration into the nature of the informal-formal sector relationship. World Dev 6(9/10):1065–1075
Tokman VE (1990) The informal sector in Latin America: fifteen years later. The informal sector revisited. OECD, Paris
Zilberfarb B-Z (1986) Estimates of the underground economy in the United States, 1930–80. IMF-Staff Papers, 33/4, pp 790–798

Further Reading

Discussions on further aspects of the informal sector can be found in the following publications:

Bajada C, Schneider F (2005) Size, causes and consequences of the underground economy: an international perspective. Ashgate Publishing Company, Aldershot
Chaudhuri S (2010) Revisiting the informal sector: a general equilibrium approach. Springer, New York/London
Thomas JJ (1992) Informal economic activity, LSE, handbooks in economics. Harvester Wheatsheaf, London

Information Deficiencies in Contract Enforcement

Ann-Sophie Vandenberghe
Rotterdam Institute of Law and Economics (RILE), Erasmus School of Law, Erasmus University Rotterdam, Rotterdam, The Netherlands

Abstract

While contracts are often useful devices for achieving commitment, they can be imperfect devices for doing so when contract breach is unverifiable by third parties or unobservable by the parties themselves. This contribution focuses on the law and economics literature which explains particular features of contract law on the basis of problems of non-verifiability and non-observability. An example is the legal system's use of weaker or no sanctions for contract breach of specific types of contracts, like employment and marriage contracts. It also includes the use of the non-verifiability problem for the evaluation of the desirability of particular legal duties, such as the duty to renegotiate contracts when circumstances change unexpectedly.

Synonyms

Non-observability; Non-verifiability

Introduction

A contract is an exchange of goods and services. However, an exchange does not necessarily require a contract. A contract is a legal instrument that puts legal pressure on the actions that parties have agreed to take at various times. Parties want their contracts to be enforced by courts to avoid the danger of opportunistic behavior. The danger of opportunism arises when performance of contractual obligations is nonsimultaneous. For then, in the absence of legal enforcement (and assuming pure self-interested behavior, no repeat play, no reputation sanction, and no taste for fairness), the last performing party has an incentive to opportunistically withhold or change his performance obligation. The problem of opportunistic behavior can be solved by drafting a contract that specifies the actions that parties are supposed to take at various times and that makes the nonperforming party subject to legal sanctions. If a party breaches the contract without a good excuse, the legal system has several choices. It can either oblige her to perform the contract as agreed (specific performance) or oblige her to pay compensation instead (damages). For example, a buyer of goods or services will often write down in a contract, sometimes in great detail, the seller's obligations. When the seller fails to satisfy these obligations, he breaches the contracts. The seller sues for damages. If the court decides that the seller did not satisfy his obligations, it will award damages. While contracts are often useful for achieving commitment, they can be imperfect devices for doing so for several reasons (Hermalin et al. 2007). First, the use of contracts to assure commitment is difficult, when due to uncertainty parties lack the information to specify the terms of their exchange in advance. This problem has been the focus of the relational contract literature (Goetz and Scott 1981). A legal requirement to force parties to specify their contracts in more detail in advance is considered not to be a good solution, as it may lead to a party's bankruptcy. Vertical integration and third-party governance have been proposed as possible solutions (Williamson 1975; Klein et al. 1978). Second, when the damages from breach are relatively low compared to the enforcement costs, enforcement may be incredible. Damage multiplicators have been proposed as a possible solution in case the probability of a suit is low (Craswell 1996).

I

But even if legal commitment has been established and the means for its enforcement are available, a contract may be an imperfect method for assuring commitment (i) when breach of a contractual obligation cannot be verified by a third party, like a judge (non-verifiability problem), and (ii) when breach is unobservable by the parties themselves (non-observability problem). There is a standard distinction in economic theory between information that parties can observe and information that is verifiable by a third party. The distinction is drawn because the costs of proving to a third party (e.g., courts) that a particular state of the world existed or a particular action was taken can exceed the gains.

Unverifiable Breach

Even if a contract party can determine that there has been a breach, she may be unable to demonstrate that fact to a third-party enforcer at reasonable cost. For example, an employer may notice that an employee who promised to work hard is shirking his duties. Still, it may be difficult for an employer to prove in court that performance was substandard. The employer cannot act on observable but non-verifiable evidence. In many cases the most important element in the success of a business is cooperative effort, which depends heavily on attitude and morale. Yet these are often the most difficult elements to explain to an outsider (Epstein 1995). The problem of non-verifiability is particularly pressing when a debtor does not commit to achieve a specific result ("obligation de résultat"), but instead commits to provide his best efforts to realize a result ("an obligation de moyens"). In the latter case, the duties undertaken are directed toward a result, but the debtor is only forced to deploy certain methods, thereby meeting the required standard of behavior, such as best efforts. Creditors often have a hard time to prove with verifiable facts that a debtor did not use his best efforts. The simple fact that the result was not achieved is not a useful proxy for low effort, since the bad result may also be due to bad luck. For example, if a lawyer is hired to work in the best interest of his client, it may be difficult to show that the quality of his pleadings is low – the fact that he lost the case may be due to factors beyond his control. The contract itself may sometimes require a debtor to prove with verifiable facts that sufficient steps were taken with a view to achieving a desirable result, but this system cannot capture all appropriate actions. Eric Posner (2000) points out that the use of expectation damages for breach of contract is problematic in cases in which the debtor's ability to perform is dependent on non-verifiable actions of the other contracting party. A contract which forces the debtor to pay expectation damages in case she fails to perform gives good incentives to the debtor to perform the contract. But suppose that the creditor is expected to take actions which make it more likely that the debtor is able to perform, but these actions are non-verifiable. Expectation damages would then give the wrong incentives because they would reward the creditor who fails to provide best efforts.

Parties sometimes do contract on the basis of observable but unverifiable information. Such contracts are, in the economic literature, said to be "implicit" or self-enforcing. They help parties to coordinate their affairs and are performed as long as coordination is mutually beneficial (or as long as reputational concerns compel compliance). An implicit self-enforcing contract is one where opportunistic behavior is prevented by the threat of termination rather than by the threat of litigation (Klein 1980). It is left to the judgment of parties concerned to determine whether or not there has been a violation of the agreement. No third party intervenes to determine whether a violation has taken place or to estimate the damages that result from such violation. If one party violates the terms, then the only recourse of the other party is to terminate the agreement after he discovers the violation. Both parties continue to adhere to an agreement if and only if each gains more from adherence to, rather than violation of, its terms (Telser 1980). People keep promises because the prospective gains from doing so exceed the prospective losses. Still, the circumstances under which the threat of termination would be a sufficient deterrent are quite severe:

the value of the contract to both parties must either be expected to continue indefinitely or have a substantial positive probability of continuing in all future states (Rosen 1984). Otherwise, there are well-known tendencies toward opportunism as the end-period approaches, and the contract is no longer self-enforcing. This problem has been discussed in economics under the name of the "last-period" problem.

Unobservable Breach

Breach is unobservable when the beneficiary of a contractual duty is unable to determine whether the promise has been kept or broken. For example, an employer may be unable to determine whether an employee has kept his promise to treat customers in a friendly manner when the employee works at a distant location. Output-based pay instead of a fixed wage may be a way to give incentives to increase the quality of the input, but is not a viable solution in case the agent is risk averse. The challenge is then to design a "mixed contract" so that there are incentives to perform well, but without burdening the agent with too much risk. The multitasking problem is a further complication of output-based pay (Holmstrom and Milgrom 1991). If payment is made dependent on the performance of one task, the agent's incentives to perform well with respect to other tasks would be severely impeded. Given the incentives and risk problems with output-based pay, many employees receive fixed wages instead. With monitoring being less than perfect, employees are to some extent free to decide how hard they will work. What motivates them to work hard when wages are fixed? What motivates them to do more than the observable and verifiable minimum standard of performance, like showing up for work during working hours? Intrinsic motivation, altruism, and an agent's identification with the principal's goals become important in this context. De Geest et al. (2001) have pointed out that the use of expectation damages for contract breach is problematic in cases in which intrinsic motivation is important for contract performance, since it may destroy such incentives.

Why the Legal System May Sometimes Use Weak or No Sanctions for Contract Breach

In general contract law, the standard damage measure for contract breach is the expectation measure. This measure awards compensation for both the reliance expenses ("negative contract interest") and the expected profits (the "positive contract interest"). The threat of having to pay expectation damages in case of contract breach is a rather heavy dose of legal pressure on contractual obligations. Still, law and economics scholarship considers it as the optimal sanction because it assures that breach only occurs when it is efficient. But for specific types of contracts, modern legal systems adopt much weaker or no sanctions for contract breach. According to the US employment-at-will doctrine, an employment relationship without a definite duration can be freely terminated. Both the employer and the employee have the right to abrogate the relationship at any time, for any reason, without notice or compensation. Parties have no ground to challenge the termination in court. The implication is that employment is "voluntary," meaning that parties perform only because they want it, not because otherwise they would be sanctioned for contract breach. Of course the absence of *legal* pressure to cooperate does not mean the absence of any pressure. *Informal* pressure may exist to trade, for example, because a party would otherwise lose a deferred benefit. Many European dismissal laws equally put little legal pressure on employment parties to cooperate with each other. In Europe, many employment relationships can be terminated without having to show just cause, provided that reasonable notice is given. Except in the limited cases where there is an abuse of rights or bad faith termination, parties have no ground to challenge the termination in court. An employee who quits early does not have to compensate the employer for the loss of expectancy, but only has to pay an amount of money in lieu of notice. Why is not more legal pressure put on employment parties to stay in their relationship? Dari-Mattiacci and De Geest (2005) explain the legal system's use of weak sanction on the basis of

the problem of non-verifiability. In order for the legal system to put legal pressure on employment parties to perform, courts would have to find out in case of employment termination which one of the parties is responsible for the failure of the employment relationship and subject that party to legal sanctions. But courts have substantial problems in verifying who was (more) responsible for the failed employment relationship. Did the employee perform poorly? Did the employer promise a more interesting and more challenging job than he offered in reality? Was one of the parties constantly unfriendly and perhaps even responsible for an unpleasant working atmosphere? These facts may be very difficult for a third party to figure out. Of course courts could easily verify which one of the parties took the initiative for the separation and sanction the initiative taker accordingly. However, if sanctions were made to depend on whether the separation is a quit or a layoff, then the distinction can quickly become blurred (Milgrom and Roberts 1992). An employer can often make an employee's life so miserable at work that he or she just has to quit, or an employee can misbehave so badly that the employer sees no choice but to fire the offender, and yet third parties cannot tell who is blamed. Such a sanctioning system will induce parties to play a "you-quit-first game." Each party will try to push the other one to quit first in order to avoid being the one who gets sanctioned. What is the outcome of this game? Who has the best chance to win a you-quit-first game?: (a) the party with the highest ability to generate negative externalities (destroying work or making life unpleasant) and (b) the party who performs poorly. If party A keeps his promises but party B breaches, staying in the relationship is least attractive for party A. As a consequence, and somewhat paradoxically, the non-breacher is most likely to quit first. In order to avoid that the "wrong" party is sanctioned for the unsuccessful employment relationship, legal systems should be reluctant to sanction contract breach when there is a low degree of verifiability. If third parties don't know who the true breacher is, it may be better not to sanction anyone in order to avoid the risk that the innocent party is the one that effectively gets the sanction. Dari-Mattiacci and De Geest (2005) also apply their framework to divorce laws, which regulate the sanctions in case of marriage contract termination. Here too the courts have difficulties in finding out which party could have done more to prevent marriage failure. Parties to a marriage contract promise to love and respect each other, but finding out which party failed first to provide best efforts to keep the obligation is very difficult. Instead, courts could use easily verifiable proxies, such as who was the party who first quit the house or started a new relationship, and sanction that party accordingly. But again this would lead to parties playing a you-quit-first game, and – as predicted by game theory – it is not necessarily the true breacher who quits first. Therefore, it may be better for the legal system not to sanction contract breach in case of divorce. A no-fault divorce system corresponds with this insight. Under no-fault divorce, neither spouse is required to prove "fault" or marital misconduct on the part of the other to obtain a divorce, and marital misconduct cannot be used to achieve a division of property favorable to the "innocent" spouse.

Should There Be a Duty to Renegotiate Contracts in Case of Unexpected Circumstances?

The 2009 Draft Common Frame of Reference (DCFR) is a European model code envisioned as a collection of "best solutions" for definitions, terminology, and substantive rules in European private law. DCFR III – 1:110 (3)(d) states that if the performance of a contractual obligation becomes so onerous because of an exceptional change of circumstances that it would be manifestly unjust to hold the debtor to the obligation, a court may vary the obligation or terminate the obligation, but only if the debtor has attempted, reasonably and in good faith, to achieve by negotiation a reasonable and equitable adjustment of the terms regulating the obligation. It has been seriously doubted whether there should be a legal duty to negotiate in good faith (De Geest 2010). There is no exact standard available to define the conduct that is required by the parties, and for courts it may be very hard to find out which party negotiated in good faith.

Because of this "non-verifiability," courts may tend to look at signals, i.e., easily observable facts that are believed to be associated with the negative behavior they want to discourage. Two such observable facts are the amount of time spent to negotiating and which party stopped negotiating first. Yet these signals may unintentionally create a you-quit-first game. Each party may try to make the other party leave the negotiation table first, so that the other one is held responsible for the failure of the negotiations. A legal duty to renegotiate is likely to cause delay and strategic behavior. Moreover, a legal duty is not necessary since contracting parties have private incentives to renegotiate their contract because in that way, they can save litigation costs.

References

Craswell R (1996) Damage multipliers in market relationships. J Leg Stud 25:463

Dari-Mattiacci G, De Geest G (2005) The filtering effect of sharing rules. J Leg Stud 34:207–237

De Geest G (2010) Specific performance, damages and unforeseen contingencies in the draft common frame of reference. In: Larouche P, Chirico F (eds) Economic analysis of the DCFR. Sellier, Munich, pp 123–132

De Geest G, Siegers J, Vandenberghe A (2001) The expectation measure, labour contracts, and the incentive to work hard. Int Rev Law Econ 21:1–21

Epstein R (1995) Simple rules for a complex world. Harvard University Press, Cambridge, MA

Goetz C, Scott R (1981) Principles of relational contracts. Va Law Rev 67:1089–1150

Hermalin B, Katz A, Craswell R (2007) Contract law. In: Polinsky A, Shavell S (eds) Handbook of law and economics, vol 1. Elsevier, Amsterdam, pp 3–138

Holmstrom B, Milgrom P (1991) Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. J Law Econ Org 7:24–52

Klein B (1980) Transaction costs determinants of "unfair" contractual arrangements. Am Econ Rev Pap Proc 70:356–362

Klein B, Crawford R, Alchian A (1978) Vertical integration, appropriable rents, and the competitive contracting process. J Law Econ 21:298–326

Milgrom P, Roberts J (1992) Economics, organization and management. Prentice-Hall, Englewood Cliffs, New Jersey

Posner E (2000) Agency models in law and economics. University of Chicago Law School, John M. Olin law and economics working paper no. 92: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=204872

Rosen S (1984) Commentary: in defense of the contract at will. Univ Chicago Law Rev 51:983–987

Telser L (1980) The theory of self-enforcing agreements. J Bus 53:27–44

Williamson O (1975) Markets and hierarchies: analysis and antitrust implications. Free Press, New York

Information Disclosure

Mika Pajarinen
The Research Institute of the Finnish Economy (ETLA), Helsinki, Finland

Abstract

This essay discusses information disclosure, i.e., making company related information accessible to interested parties, in regard to economic literature. The two main fields in which information disclosure is dealt with in this context are corporate finance and innovations. Short notes on other contexts in economics are also made. The essay excludes the descriptions of formal procedures and legal aspects of disclosing information.

Definition

To make (company related) information accessible to interested parties.

Information Disclosure in Corporate Finance

In corporate finance literature information disclosure relates closely to investment and growth opportunities. Firms whose financing needs exceed their internal resources may suffer from financial market imperfections due to asymmetric information and incentive problems between corporate insiders and outside investors (see, e.g., Berger and Udell 1998; Hubbart 1998; Petersen and Rajan 1994). The informational opacity of a firm may reduce the availability of external finance to the firm because the more opaque the firm the more scope there is for opportunistic

behavior by the firm's insiders (more moral hazard) and the harder it is for investors to determine the quality of the firm (more adverse selection). This may lead to higher interest rates demanded by investors and higher risk investment projects chosen by a firm than in a situation in which all relevant information on firm's financial status has been made accessible to potential investors. A conventional wisdom in the contemporary corporate finance literature argues the informational opacity of firms is in relation to firms' age and size. Recent entrants with short track record suffer more from the informational opacity than incumbent firms. Smaller firms are also typically more opaque than larger firms because their minimum requirements for information disclosure, e.g., in financial statements, are more scant than in larger, especially in the listed, firms.

Firms can reduce their informational opacity by voluntarily disclosing high quality information on their business activities over and above mandated disclosure. High quality disclosure is especially important for firms with lucrative growth prospects because for them standard disclosure is of too low quality. It is of too low quality in the sense that mandated disclosure often alleviates information asymmetry only to a limited extent. This kind of higher quality disclosure incurs, however, costs to firms. There can be many types of costs. Direct costs arise from producing and credibly disclosing information. In addition to more comprehensive accounting processes, this may include, for instance, acquiring prestigious auditors and the premiums charged by them. Furthermore, indirect costs can arise from various reasons, such as the pro-competitive effects of disclosure (Bhattacharya and Chiesa 1995; Healy and Palepu 2001). Disclosure may also reduce the incentives of outside investors to acquire information (Boot and Thakor 2001). Firms have thus to balance with the costs of high quality disclosure with returns of growth opportunities. Titman and Trueman (1986) show in a formal model that the firms that choose high quality disclosure are, in equilibrium, those with favorable information about the firm's future and its growth opportunities.

Information disclosure in corporate finance is related also to the interactions of the insiders of firms, i.e., corporate governance system. Hermalin and Weisbach (2012) argue that more transparent disclosure policies can in fact aggravate agency problems between managers and shareholders and increase firms' costs due to higher rates of executives' turnover and compensation demanded by them. Efforts to be more transparent can lead to a situation where managers are more reluctant to increase firm value in the long run but rather boost short-term investments and other actions that affect reported figures sooner at the expense of longer term (riskier) value creation investments, such as R&D.

Information Disclosure and Innovations

Besides corporate finance, information disclosure relates in economic literature also to generation of innovations. When a firm has made an innovation, it can basically choose to keep the innovation secret or apply for a patent. If the firm applies for the patent, it discloses the discovery of invention and provides information about it to the public. In addition, the patent applicant is obliged to reveal all prior art that may be relevant to the patentability of the applicant's invention. This includes disclosing existing technological information on both documentary sources, such as patents and publications, and nondocumentary sources such as things known or used publicly, which is used to determine if an invention is novel and involves an inventive step (for further details on patenting process, see, e.g., European Patent Office's www-pages http://www.epo.org).

When a firm is granted a patent on an invention, it gets a temporary monopoly right in exchange for disclosure. The monopoly right lasts normally 10 years. After the period of patent protection, the invention is freely available to competitors and other potential users. Other parties than the inventor can also utilize the patented innovation during the life of the patent by licensing and other arrangement facilitating a market for technological exchange. Social costs of disclosing the invention by patenting arise if this technological exchange does not perform well. In this case, patented inventions may hold up subsequent research on related inventions and

may generate substantial transaction costs from costly legal challenges about possible infringement (see Gallini 2007 for a more detailed discussion of the role of disclosure in the case of patenting and Griliches 1990 for a review of patents as innovation and economic indicators).

Information Disclosure in Other Contexts

Besides corporate finance and innovations, the term information disclosure is used for instance in consumer economics in regard to how transparent information firms are willing to reveal on their products to consumers (see, e.g., Ghosh and Galbreth 2013; Polinsky and Shavell 2012).

Furthermore, in the field of competition analysis, information disclosure is used in the context of its influence on competition in the field of interest (see, e.g., Arya and Mittendorf 2013; Feltham et al. 1992; Hayes and Lundholm 1996). More detailed information disclosure reveals more data to competitors and can potentially weaken disclosing firm's position in the market. On the other hand, an incumbent firm may strategically choose to disclose some information to prevent new entries in the market. In addition to competing firms, individual firm's disclosures may have spillover effects also in non-competing firms by revealing information on technological trends, governance arrangements, best policy practices, etc.

Cross-References

- ▶ Auditing
- ▶ Externalities
- ▶ Innovation

References

Arya A, Mittendorf B (2013) Discretionary disclosure in the presence of dual distribution channels. J Account Econ 55:168–182

Bhattacharya S, Chiesa G (1995) Proprietary information, financial intermediation, and research incentives. J Financ Intermed 4:328–357

Berger A, Udell G (1998) The economics of small business finance: the roles of private equity and debt markets in the financial growth cycle. J Bank Financ 22:613–673

Boot A, Thakor A (2001) The many faces of information disclosure. Rev Financ Stud 14:1021–1057

Feltham G, Gigler F, Hughes J (1992) The effects of line-of-business reporting on competition in oligopoly settings. Contemp Account Res 9:1–23

Gallini N (2007) The economics of patents: lessons from recent U.S. patent reform. In: Economics of intellectual property law. volume 1. R. P. Merges, Elgar Reference Collection, 16. Elgar, Cheltenham/Northampton, pp 63–86

Ghosh B, Galbreth M (2013) The impact of consumer attentiveness and search costs on firm quality disclosure: a competitive analysis. Manag Sci 59:2604–2621

Griliches Z (1990) Patent statistics as economic indicators: a survey. J Econ Lit 28:1661–1707

Hayes R, Lundholm R (1996) Segment reporting to the capital market in the presence of a competitor. J Account Res 34:261–279

Healy P, Palepu K (2001) Information asymmetry, corporate disclosure, and the capital markets: a review of the empirical disclosure literature. J Account Econ 31:405–440

Hermalin B, Weisbach M (2012) Information Disclosure and Corporate Governance. J Financ 67:195–233

Hubbart G (1998) Capital-market imperfections and investment. J Econ Lit 36:193–225

Petersen M, Rajan R (1994) The benefits of firm-creditor relationships: evidence from small business data. J Financ 49:3–37

Polinsky A, Shavell S (2012) Mandatory versus voluntary disclosure of product risks. J Law, Econ Org 28:360–379

Titman S, Trueman B (1986) Information quality and the valuation of new issues. J Account Econ 8:159–172

Information Privacy

▶ Privacy

Innovation

Tanja Benedict
InnovationLab GmbH, Legal Services and IP Management, Heidelberg, Germany

Synonyms

Creativity; Novelty

Definition

The process of devising a new idea or thing or improving an existing idea or thing (Sandefur 2008). Innovation is also defined as "the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations." It differentiates between four types of innovations, namely, "product Innovation," "process Innovation," "marketing Innovation," and "organizational Innovation" (OSLO Manual, OECD 2007).

The Term Innovation

Meaning of the Term

The introduction of something new; a new idea, device, or method (Latin: innovatio: renovation, replacement, change, novelty). Innovation in a general sense is understood as renovation and redesign of divisions and parts of economy, science or society, in specific functioning models or behavior patterns in relation to an existing system (economically or socially). Innovation is understood to be an improvement of existing procedures, materials, or devices or a better match to changed requirements of functionality. Innovation has to be distinguished from invention and diffusion. Invention focuses on the mental process of creating a new idea. The emphasis is lying in the moment of creation, the birth of something, that is brought into the material world. In contrast to the term invention, innovation requires also the aspect of the new idea being applied or performed in a way that matters within the context of the innovation. An innovation has to introduce relevant changes on the applied level. Not every invention is necessarily an innovation. In comparison to sustaining innovations, innovations with a high relevance of change are called *disruptive innovations* (see "Clayton M. Christensen: Disruptive Innovation"). Diffusion describes the following process of spreading and transferring the applied invention in the market. Historically the term has been used in French since the thirteenth century in its general meaning, in English it was recorded since the fifteenth century. Up to the twentieth century, the term had a specific meaning merely in botanic sciences and legal procedures. After World War II, the term has spread through reception of macroeconomic theories on the international level.

Innovation as a Term of Economics

Innovation as a term of economics has been introduced by the Austrian-American economist *Joseph Schumpeter* in his *Theory of innovations* first published in 1912. Innovation as a concept of implementation of new products, processes, or management ideas is strongly connected to Schumpeter's idea of creative destruction. Creative destruction occurs, when innovations make existing products and services obsolete, thus freeing resources to be employed and used elsewhere. So, creative destruction leads to greater efficiency. The creative entrepreneur profits from return of investment by temporary monopoly through innovation instead of profiting from exploitation of price differences. According to Schumpeter, the process of innovation consists of three parts: invention (the creation of the new idea), the innovation itself (implementation of the invention), and the market diffusion (production and sale). A product or process is called innovative only if there is market diffusion, that makes the difference between the mere invention and the innovation.

Innovation as a Legal Term

Innovation is not regarded a technical legal term. Until 2014 there has not been a legal definition of innovation, although the frequency of the term in statutory European and national law is steadily increasing. European Directive 2014/24/EU on public procurement defines innovation in Art. 2 (22) as "the implementation of a new or significantly improved product, service or process, including but not limited to production, building or construction processes, a new marketing method, or a new organizational method in business practices, workplace organization or external relations inter alia with the purpose of helping to solve societal challenges or to support the Europe 2020 strategy for smart, sustainable and inclusive growth;"

The term innovation has been introduced by the European Commission already in 2010 within the programmatic term *Innovation Union*, which is part of the strategy Europe 2020. Innovation in this sense means more than being innovative. "The EU initiative Innovation Union focuses Europe's efforts – and its cooperation with non-EU countries – on the big challenges of our time: energy, food security, climate change and our ageing population. It uses public sector intervention to stimulate the private sector and remove bottlenecks which prevent ideas from reaching the market – including lack of finance, fragmented research systems and markets, under-use of public procurement for innovation and slow standard-setting." (Communication of the Commission 2010) Here innovation is more of an action program following specific political aims. Whereas innovation as an economic concept describes the economic success of innovative products through efficiency in some relation.

Innovation as a Term in Social Sciences

What Schumpeter developed for economics, Friedrich August von Hayek developed in the political theory of the "Open Society," where he stressed the importance of innovation for social values. Hayek believes in the existence of an order in society. "It would be no exaggeration to say that social theory begins with - and has an object only because of - the discovery that there exist orderly structures which are the product of the action of many men but are not the result of human design" (Hayek 1944). He assumes that planned orders are inferior and will not produce prosperity. Still prosperity requires some kind of regulation by man, as the rule of law, to prevent innovation being steered in a parasitic direction. "Natural selection operates on mutations, making the path of natural selection unpredictable, regardless of how well we understand the underlying principles" (Hayek 1944). To Hayek, social and cultural evolutions are much the same: driven by innovation, fashion, and various shocks that "mutate" people's plans in unpredictable ways with unpredictable results. The system may be more or less logical. And by no means predictable.

Concept of Innovation in Economics

Joseph Schumpeter: Creative Destruction

Schumpeter defines innovation as a process of changes focused on the creation of something new. The combination of various economic factors in a new, "innovative" way has an impact on the economy and the society on the whole. According to Schumpeter, the process of technological change has three levels: invention (the creation of a new idea), innovation (setting up the elements for the implementation of the invention), and market diffusion. Innovation has an overall positive connotation. But as every human activity innovation has costs as well as benefits. Instead of bringing relieve, creating wealth and power, innovations can also disrupt the status quo. Schumpeter calls the process of creative destruction the process by which innovation causes resources to be set free and therefore introduced the term creative destruction. Creative destruction occurs when innovation makes production processes obsolete and frees resources that can be employed elsewhere, thus increasing efficiency. Creative destruction has always been feared as source of unemployment. The term creative destruction is sometimes also known as Schumpeter's gale, which Schumpeter derived from Karl Marx and integrated it into his theory of economic innovation. According to Schumpeter, the "gale of creative destruction" describes the "process of industrial mutation that incessantly revolutionizes the economic structure from within, incessantly destroying the old one, incessantly creating a new one" (Schumpeter 1939). In the earlier work of Marx, the idea of creative destruction or annihilation implies not only that capitalism destroys and reconfigures previous economic orders. It must also ceaselessly devalue existing wealth (e.g., through war, dereliction or economic crises) in order to clear the ground for the creation of new wealth. Schumpeter believes in the existence of an economic equilibrium, whereas Marx defines the boom as a consequence of a depression. The motor of innovation according to Schumpeter is the position of monopoly, which the creative entrepreneur will seize by introducing an innovation.

Clayton M. Christensen: Disruptive Innovation

Not every innovation is of the same kind: there are sustaining innovations made by incremental research, aiming at keeping up technologically with competing firms. Most inventions produce sustaining innovation. Disruptive innovation describes a process, where innovation creates a new market and value network that will eventually disrupt an already existing market and replace existing products. It is more the business model that creates the disruptive impact, than the existence of one high technology invention. Disruptive innovation refers rather to the evolution of a product or service than to a product or service at a certain point. Harvard Business School economist Clayton M. Christensen introduced the term disruptive technologies in his article *Disruptive Technologies: Catching the Wave* (1995) with his cowriter Joseph Bower and described the term further in *The Innovator's Dilemma*. Later he replaced the term by disruptive innovation when he recognized that few technologies are intrinsically disruptive or sustaining, but that it is rather the business model enabled by the technology that has a disruptive character. The evolution from a technological focus to a business modeling focus is a central part of his concept.

Christensen said: "The technological changes that damage established companies are usually not radically new or difficult from a technological point of view. They do, however, have two important characteristics: First, they typically present a different package of performance attributes-ones that, at least at the outset, are not valued by existing customers. Second, the performance attributes that existing customers do value improve at such a rapid rate that the new technology can later invade those established markets" (Christensen 1995). Christensen differentiates between low-end disruption and new-market disruption. The first is targeting at customers that do not need the full performance value; the second is targeting at new customer groups whose needs have not been served by existing incumbents. Christensen's theory of disruptive innovation had a strong influence on business and his book was chosen for one of the most important books in economics. Eventually the term disruptive innovation was said to be overused and to have become a cliché among people, partly because they do not understand Christensen's concept.

Functionality of Law for Innovation

Innovation as a process is not intrinsically a legal matter. But innovation requires an atmosphere, where values and rights indispensable for innovation have to be protected by the law. Law may be used, to protect existing values in a society, but not to create them. In the same time law has by means of regulations to avoid "parasitic" directions to mention Hayek, if innovation shall create prosperity. "A primary role of law and (when necessary) legislation is to narrow people's options so as to limit opportunities to get rich at other people's expense. So long as the rule of law can internalize external cost and thereby steer innovation in mutually beneficial rather than parasitic directions, an evolving order will be an order of rising prosperity." Law therefore has at least two functions for innovation: protecting the basis for its requirements and steering it by regulation. Law can open up the field for innovation, support enabling innovators to act, set incentives, steer and regulate. The functionality of law therefore is more of an instrument. The innovation process describes different stages of economic development that require legal regulation in various areas of law. As innovation regularly implies inventions or creation of know-how innovation is above all strongly connected to the law of intellectual property. To allow something new to emerge and to keep the position of monopoly created by the implementation of an invention, innovation requires the legal regulation of rights that embody inventions or know-how and can be used as tradable carrier of that right. So, intellectual property law can be regarded as the legal core of innovation (but not the basis). Neighboring subjects as the law of employees' invention and contract law (licensing contracts, research & development contracts) contribute significantly to innovation too. Innovation requires confidentiality, protection of trade secrets, exclusivity and protection against competitors. But these are only

legal topics on the surface of innovation. Innovation requires not only new ideas but also entrepreneurship and capital. In legal terms, innovation therefore requires basic rights and freedoms, as the freedom to do research and to work in an open academic environment with scientific methods. Capital needs to be stable and mobile.

But law may also be used to trigger or facilitate innovation. So legal regulations on funding R & D, enabling public-private partnerships in R & D, company law that offers easy to handle formats for companies and facilitates fast registering, tax law privileging insurance companies offering risk capital to start ups, labor law regulations that enable flexibility in hiring people, or low rates of courts and offices in patent matters may have strong impact on innovation. Innovation is not one legal field, but rather a cross-cutting issue.

Legal Basis of Innovation - the Legal Side of Entrepreneurship

Entrepreneurship is understood as the ability and willingness to act creative. The ability to act creative requires a spirit of individualism and the development of scientific methods in a society. The law can support this spirit and capacities by protecting the basic rights and freedoms: freedom of scientific research, freedom of movement, freedom of establishment, and others. Innovation needs to create its own requirements. Innovation may be a rule of interpretation on the level of constitutional law when it comes to fundamental understanding and interpretation of laws and jurisdiction. The relation between innovation and stability has to be solved in favor of the development of constitutional law, rather than preservation and persistence. In federal states, procedural means to compensate differences in federal judiciary by examining decisions on the level of constitutional law regarding their congruence, as well as the introduction of actions for natural persons on the constitutional level are also ways to open law to innovation. Freedom in business means to provide for competition. In Hayek's words, "the law should not only recognize the principle of private property and freedom of contract, but the legal system should also give a precise definition of these two principles in a way that promotes competition" (Hayek 1944). With the fast growth of high-tech markets the phenomenon of monopolies and dominant market positions have to be observed closely and if necessary regulated to conserve competition. But innovation has changed structures of markets too: "market shares move faster, barriers to entry the market tend to be much lower, and natural monopolies leave as fast as they come" (Schrepel 2014, 216). Antitrust law therefore has to be updated to these new market mechanisms. "A paradox clearly appears on this point. Indeed, very few persons argue that antitrust should promote perfect competition. However, not to take every single aspects of innovation into account, – for instance, by not including disruptive and permissionless innovations in our analyses – has for consequence to indirectly promote this model of perfect competition. For the reason, it is time to hold innovation as real antitrust standard "(Schrepel 2014, 216).

Confidentiality

Innovation exploits the fact that a product, material, service, or business model is new to the rest of the market and therefore enables a monopoly. Once the invention is to become an innovation, it has to be filed as a patent and implemented in the market as a product. To uphold the monopoly as long as possible, the entrepreneur needs to protect himself against competitors by a ban to use his invention, a patent, by keeping most of his technological competence secret, by defining know-how as trade secret of his business. In all cases, you need confidentiality, which usually is created by non-disclosure agreements, material transfer agreements, and obligations of confidentiality in labor contracts.

Confidentiality has to be limited in time and relevance. An invention kept secret absolutely can hardly have an innovative impact. Sometimes the position of a monopolist can be upheld by suppressing innovation that could disrupt the incumbent product. On the long run though, it is difficult to suppress innovation entirely. So on the

one hand confidentiality is required to create monopoly; on the other hand, there will be no profit without implementation, which means the invention has to be made public to some extent. As innovation implies inventions and invention imply novelty, you need a protection of your research and development results before you file them as a patent. As a result non-disclosure agreements regulate a graded confidentiality, limited to the relevant topics, limited in time and people, who need to know. They oblige their partners to keep information defined as confidential secret and not to use it in any other business relation. Limitation of the obligation here is fundamental, without it, entrepreneurs may end up in an information blockade that eventually destroys business opportunities. Attempts to regulate confidentiality on a European level have not been successful. As non-disclosure agreements are used to create trustful relations between partners, it might be better not to make them obsolete by legal regulation.

Intellectual Property Law

Intellectual property law describes legal rights designed to enable technology transfer and doing business with IP. Patents, utility models, designs, registered marks, as well as trade secrets and know-how are means to protect the innovation against competitors and to keep the monopolistic position as long as possible. The patent is used for the protection of inventions. In the global market, it was very important to have a common understanding and a similar regulation of the patent in the different countries. With the Patent Cooperation Treaty, the Paris Convention for the protection of Industrial property rights, the foundation of the World Intellectual Property Organization there is a basis for economical exploitation of inventions in the global market. A patent is the right to forbid anyone to use the invention that is filed for a patent in the geographical area of the patent for a maximum of 20 years. Global Players need to register in all countries, where they intend to use the patent, which is, where they produce and sell a product, based on the patent. This can result in high costs, depending on the number of countries the patent should be registered for. The patent is the strongest intellectual property right, especially compared to know-how or utility models. An essential difference between regulations on patents between European and US law is the patentability of business models in US Law, whereas under European Law business models can only be protected by means of protecting its designs or registered marks used by the business model. In Europe, a patent can only be granted for the solution of technical problems. Considering that the so-called platform businesses profit more from their business model, than from a specific high technology invention, the scope of application of European patents should be rethought. The European initiative to introduce a common European Patent and a corresponding judiciary has been stopped for the time being – not by Brexit negotiations, but by court actions challenging the judicial independence of European Patent authorities, currently discussed vigorously.

Law as Stimulation and Facilitation

Law can stimulate innovation by facilitation of legal processes or introduction of completely new legal forms. For example, the Innovation Partnership, a special form of procurement procedure in European public procurement law, introduced in 2014, is a specific procurement procedure with the aim of developing an innovative product or service and then acquiring the resulting services. Due to the special European procurement law for the member states, the legal institution also works in its procurement law which enables flexibility within the European market. A characteristic of the procedure is a respective two-stage of award procedure and the subsequent contract model. Many public procurement regulations on European level as well as on national levels are directed to the promotion of innovation.

Legal Requirements during the Process of Innovation

The process of innovation usually describes the development starting with research and

ending with the revenue gained after commercialization of the innovative product, service, or business model. Depending on the affiliation of authors, the process of innovation may concentrate on the first part and be called Technology Transfer process or concentrate on the commercialization and start up business. Among the plenty and various models for the innovation process, most of them can be divided into three larger phases, subdivided by different steps.

Phase 1: Conception/invention/knowledge phase: This includes idea generation and evaluation, pre-disclosure, and invention disclosure assessment protection. On the legal side this requires potentially non-disclosure agreements, labor contracts for scientists with paragraphs on intellectual property and confidentiality, invention disclosures, patent or utility filings, patent investigation, and more.

Phase 2: In the second phase, the implementation or adaption phase, the invention made in phase 1 is developed into a product, a prototype, or even more a pilot series, ready for being tested. On the legal side this is accompanied by product development agreements with the company that intend to develop a product. This can either be an existing company or a start-up. In the latter case, the process is completed by financing. Usually the different processes overlap. Once a pilot series has been concluded successfully, innovation is ready for production in phase 3.

Phase 3: It starts with the licensee agreement, production, and continues with the market launch, eventually aiming at market penetration. Legally the step from phase 2 to 3 is enormous, because the law treats production very different from research and development. Liability of manufacturer, quality management, technical data sheets, higher insurance rates, different general conditions are only some of the legal aspects, production and sale require. This is a reason, why usually production is located in a different legal entity to keep the higher risks separate from research and development.

Impact of Innovation on the Development of Law. New Legal Questions?

Law can influence innovation, as well as innovation has an impact on the development of law. Digitalization, artificial intelligence, and information technologies let us rediscover old problems in a new form, but sometimes also raise new questions. The relation between these innovative technological fields and law, the question of regulation of technology is a cross-section field: liability, traffic law, data protection law, and criminal law. Digitalization requires data protection and raises the question of information privacy, especially when digitalization is introduced in the public sector for e-government, as introduced by the European Union with the eGovernment action plan 2016–2020 and e-government laws on the national level as the German law in 2013. As mentioned before, on one hand law will have to facilitate innovation by enabling mobility of citizens and businesses by cross-border and facilitating digital interaction between administrations and citizens/businesses for high-quality public services. On the other hand, it will have to limit the amount of personal data to be collected and protect citizens against misuse of personal data.

Data protection law in Europe is criticized for fulfilling its purpose only imperfectly. In wide ranges, it is still based on the data protection concept of the 1970s, when central data storage on mainframes, limited storage capacities, and a relatively small circle of – mostly government – data processors were characteristic. Only slowly legislation can keep up with the needs of technical development. The European Union started a general data protection reform and issued the General Data Protection Regulation in 2016, according to which the Data Protection Directive of 1995 expires. On the national level, data protection law, e.g., in Germany, is considered "over-regulated, fragmented, confusing and contradictory" (Roßnagel et al. 2001).

Digitalization requires the regulation of issues raised by the increasing use of automated and autonomous devices. Especially the fact that we

allow machines to make decisions for us raises significant problems of liability, responsibility, and the nature of legal personality. For the self-driving car law, for example, we will have to decide, which risks will be carried by manufacturer's responsibility and what will lie within the driver's liability. The solution of dilemmas in emergency situations, to be decided by machines, may not be solvable by law, but rather by ethics, sociology of technics, and "interdisciplinary methods to be developed" (Valentiner 2016).

Summary

Innovation is the process of devising a new idea or thing or improving an existing idea or thing. In contrast to the term invention, innovation requires also the aspect of the new idea being applied or performed in a way that matters within the context of the innovation. Innovation as a term of economics has been introduced by *Joseph Schumpeter* and is strongly connected to his idea of creative destruction. Clayton Christensen coined the term destructive innovation. His theories had a strong influence on business administration. Innovation requires entrepreneurship and capital. The function of law is to protect the basis for innovation and to regulate and steer it. Digitalization, information technologies, and artificial intelligence are challenges to the law. After decades of rising numbers of legal regulations, innovation could take us to the limits of law.

Cross-References

- ▶ Creativity
- ▶ Entrepreneurship
- ▶ European Patent System
- ▶ Hayek, Friedrich August von
- ▶ Information Disclosure
- ▶ Patent Litigation
- ▶ Patent Opposition
- ▶ Rule of Law
- ▶ Trade Secrets Law

References

Christensen Clayton (1995) Disruptive technologies. Catching the wave, Harvard business Review, January

Communication of the Commission Mitt. der Kom. v. 2010 (KOM) (2010) „Leitinitiative der Strategie Europa 2020 Innovationsunion"

Hayek F (1944) Road to serfdom, 2001 first published 1944. Routledge Classics, London

Roßnagel Alexander, Pfitzmann Andreas, Garstka Hansjürgen (2001) Modernisierung des Datenschutzrechts. Gutachten im Auftrag des Bundesministeriums des Innern. Berlin

Sandefur T (2008) Innovation. In: The concise encyclopedia of economics. Springer, New York

Schumpeter JA (1939) Business cycles. A theoretical, historical, and statistical analysis of the capitalist process. MacGraw-Hill, New York

Valentiner Dana (2016) Gesetzgeberische Herausforderungen der Technikregulierung – ein Aufriss, juwiss.de/43–2016/

Further Reading

Christensen Clayton, Raynor Michael, McDonald Rory (2015) What is disruptive innovation? Harvard Business Review, Dec 2015

Djeffal Christian (2017) Leitlinien der Verwaltungsnovation und das Internet der Dinge, Alexander von Humboldt, Institut für Internet und Gesellschaft

Fritsch M, Werker C (1999) Innovation Systems in Transition. Innovation and Technological Change in Eastern Europe: pathways to industrial recovery. M. Fritsch and H. Brezinski. Cheltenham, UK, Edward Elgar

Hayek F (1944) Road to serfdom. Routledge Press, London

Hoffmann-Riem, Wolfgang (2016) Innovation und Recht – Recht und Innovation, Recht im Ensemble seiner Kontexte

Kaschny M, Nolden M, Schreuder S (2015) Innovation management in SMEs: strategies, implementation, practical examples. Springer Gabler, Wiesbaden

Kühling J, Seidel C (2015) Anastasios Sivridis: Datenschutzrecht, 3rd edn. Müller, Heidelberg

Lepore Jill (2014) The disruption machine, The New Yorker. June 23

Schmidtz David (2016) Friedrich Hayek. In Edward N. Zalta (ed) The Stanford encyclopedia of philosophy (Winter 2016 edn), URL = https://plato.stanford.edu/archives/win2016/entries/friedrich-hayek/

Schrepel T (2014) Friedrich Hayek's contribution to antitrust law and its modern application. Global Antitrust Review:199–216

Schulz Wolfgang, Dankert Kevin (2016) Governance by things as a challenge to regulation by law, IPR 5

Thurston Thomas, Crunch Network (2014) Christensen Vs. Lepore: a matter of fact, posted 30 Jun 2014 by Thomas Thurston (@thurstont)

Tinnefeld Marie-Theres, Buchner Benedikt, Petri Thomas (2012) Introduction to data protection law. In: Data

protection and freedom of information in a European perspective. 5th edn. Oldenbourg
von Hippel Eric The sources of innovation, Oxford University Press, New York 1988, Download courtesy of OUP at http://web.mit.edu/evhippel/www

Institution

► Organization

Institutional Change

Christopher Kingston
Amherst College, Amherst, MA, USA

Definition

Institutions are durable; that is precisely what makes them meaningful and important. But institutions also sometimes change. This entry compares a variety of theoretical approaches to understanding the process of institutional change. Some authors treat institutional change as a centralized, collective-choice process in which rules are explicitly specified by a collective political entity, such as the community or "the state," and individuals and organizations engage in collective action, conflict, and bargaining to try to change these rules for their own benefit. Others emphasize the "spontaneous" emergence of institutions as an evolutionary process, in which new institutional forms periodically emerge and undergo some kind of decentralized selection process as they compete against alternative institutions. Still others combine elements of evolution and design. We differentiate a variety of approaches to the interaction between formal and informal rules and explore the path-dependent nature of institutional change. We also discuss recent theories based on the "equilibrium view" of institutions, which emphasizes that the constraints that motivate individual behavior are ultimately derived from expectations about the behavior of other actors in various contingencies. In maximizing their welfare subject to these constraints, agents choose strategies which, in the aggregate, give rise to expectations which reinforce the constraints on everyone else, so the effective "rules" of the game emerge endogenously as equilibrium outcomes, and exploring institutional change involves explaining changes in equilibrium behavior.

Theorizing Institutional Change

The scholarly literature on institutional change is voluminous, but diffuse and eclectic, plagued by ambiguity about the meaning of commonly used terms including even the meaning of the term "institutions" itself. In this entry, adapted from a broader paper on the subject (Kingston and Caballero 2009), I attempt to map out the major theoretical approaches to institutional change and to highlight the main possibilities that an empirical researcher might consider.

Before we can discuss institutional change, we must define what we mean by "institutions." Unfortunately, the appropriate definition is far from a settled issue, and the different definitions used by different authors naturally influence their views of institutional change. Many authors, however, adopt some variant of Douglass North's view that institutions "are the rules of the game in a society or more formally, are the humanly devised constraints that shape human interaction" and that they "reduce uncertainty by providing a structure to everyday life" (North 1990: 3). Fundamentally, then, institutions are viewed as "rules" that are "humanly devised."

There are many different kinds of "rules"; however, most authors follow North in distinguishing between "formal" rules such as laws and constitutions and "informal" constraints such as conventions and norms. Even this basic distinction is far from straightforward. The term "formal" is often taken to mean that the rules are made explicit or written down, particularly if they are enforced by the state, whereas "informal" rules are implicit; another interpretation is that formal rules are enforced by actors with specialized roles, whereas informal codes of behavior are

enforced collectively by the members of the relevant group. "Informal constraints," as North notes, "defy, for the most part, neat specification" (North 1990: 36) but include "socially sanctioned norms of behavior" as well as "extensions, elaborations, and modifications of formal rules" and "internally enforced standards of conduct" (North 1990: 40).

But while social norms enforced by a community, conventions, and internalized ethical codes such as religious beliefs can all be viewed as informal constraints, they are distinct phenomena which may change in different ways and may have different short-run and long-run effects on the broader pattern of institutional change. This ambiguity has created considerable confusion about the nature of informal constraints and their interaction with formal rules. For example, some authors have regarded informal constraints as essentially immutable, or that they change only slowly, but there have also been episodes when at least some kinds of informal constraints, such as social norms, can change rapidly.

For the moment, then, let us treat institutions as the (formal and informal) "rules of the game" in a society; we will discuss an alternative definition later. How do these rules change? Two main kinds of processes emerge from the literature. Some authors view institutional change as the outcome of purposeful (centralized) design, either by a single individual (such as when a king issues a royal decree) or by many individuals or groups interacting to create or change rules through some kind of collective-choice or political process. Other authors envision institutional change as a more gradual, evolutionary (decentralized) process, frequently involving competition among alternative institutional forms. In empirical settings, aspects of both kinds of processes are often present, and the question becomes how they interact. This raises a host of conceptual issues, including how we should think about the interaction between formal rules and informal constraints, the role of politics and collective action, the nature of "competition" between different institutional forms, the role of bounded rationality and learning, the exogenous and endogenous causes of institutional change, and the role of history and the potential for path dependence.

Politics and Collective Choice

Many authors treat institutional change as a centralized, collective-choice process in which rules are explicitly specified by a collective political entity, such as the community or the state, and individuals and organizations engage in collective action, conflict, and bargaining to try to change these rules for their own benefit.

Ostrom (2005), for example, envisions a multilayer nested hierarchy of rules: "operational rules" that govern day-to-day interactions, "collective-choice rules" (rules for choosing operational rules), and "constitutional rules" (rules for choosing collective-choice rules). In order to analyze how rules are formed at one level, Ostrom temporarily treats the higher levels of rules as fixed (ibid.: 61). For example, when "operational rules" are being chosen, constitutional and collective-choice rules are treated for the moment as exogenous.

The impetus for institutional change arises when some group or individual perceives an opportunity to change rules in a way that benefits them. This could be because of an exogenous change in underlying parameters that change the perceived costs and benefits from an institutional change – for example, a change in technology or relative prices. But the cause might also be endogenous if, say, people's choices under one set of rules gradually lead to changes in parameter values that alter the costs and benefits of institutional change and thereby undermine current institutions. For example, Acemoglu and Robinson (2005) argue that after 1500, some European countries experienced a substantial growth in Atlantic trade which increased the political power of merchant groups. The growing strength of merchant groups in these countries led to institutional changes that constrained the power of monarchs and led to the development of institutions that were more conducive to economic growth.

The process of institutional change, in Ostrom's framework, is this: each individual calculates their expected costs and benefits from an institutional change, and if a "minimum coalition" necessary to effect change agrees to it, the change is enacted. What constitutes a "minimum coalition" is determined by the higher-level rules; for example, in a dictatorship, the dictator alone might constitute a winning coalition; in a democracy, a majority would constitute a winning coalition. The outcome therefore depends on how decision-makers perceive the likely effects of a change in rules and on whether those that desire change are able to bring it about or whether those that expect to lose by the change are able to block it under the higher-level rules which frame the political (rule-making) contest. Powerful groups may be able to block beneficial change or impose inefficient change. Of course, some kind of compromise or partial institutional change is also a possible outcome.

Free-rider problems can impede collective action to change formal rules. Voting, protesting, joining political associations, and learning about the impact of potential policies may all be individually nonrational actions, even if the individual cares deeply about the result. Leaders can play a key role by offering a "vision": attempting to shift their supporters' perceptions of the costs of the existing system or the feasibility and desirability of some proposed alternative. Whether they seek wealth or power, or are driven by ethical or philosophical values, the fundamental challenge that leaders must overcome in order to achieve institutional change is that of winning support and overcoming collective action problems among their potential supporters.

The role of political actors, such as judges and politicians, can be envisaged in a variety of ways. For some purposes, politicians might be viewed as simply reflecting the interests of particular groups, so that the political process remains essentially a battleground in which interest groups compete to mold formal rules to their own advantage. Other theories, however, give political actors a more autonomous role. For example, in Kantor's (1998) framework, groups of constituents lobby politicians to change formal rules, and the politicians have incentives to be responsive to their constituents' demands. However, the politicians also have their own objectives and face other political and constitutional constraints. Which of these perspectives on the role of political actors is most appropriate will depend on the configuration of interest groups and the political structure in a given context.

Another important set of barriers to institutional change arises when the beneficiaries of a reform cannot credibly commit to compensate the losers because of the lack of an external authority to enforce intertemporal bargains. These problems are exacerbated when there is uncertainty or when there are a large number of parties to a negotiation. Acemoglu and Robinson (2006) highlight the importance of commitment problems as a cause of institutional change. In their theory, disenfranchised groups can use violence to force constitutional change (seize power), but because of collective action problems, they can only organize violence during rare (and exogenous) moments of crisis. Violence is destructive, so in a crisis, there is an opportunity for a mutually beneficial bargain in which the incumbent ruling groups would agree to carry out reforms, in exchange for which the disenfranchised groups would refrain from carrying out a revolution. However, the disenfranchised groups' opportunity to use force is fleeting, and the ruling groups cannot credibly commit themselves to honor their commitments to reform after the moment of crisis is passed. Therefore, a revolution to effect institutional change, though destructive and costly, may be the only way for currently disenfranchised groups to credibly constrain the policy choices of future governments.

Many authors note that institutional change a "path-dependent" process – that is, the institutions observed at any point in time, may in part be a function not just of current technology but also of precedent institutions and technologies (David 1994, 2007). Frequently, existing institutions create groups with a vested interest in preserving the status quo, which can impede institutional change and enable inefficient institutions to persist. More generally, the resources, physical and human capital, skills, technologies, and

organizations accumulated under one set of institutions can gradually alter the set of technologically feasible institutions and thereby affect future institutional development. Furthermore, even when these impediments are overcome, institutional change is usually incremental since it is often easier to achieve consensus on small adjustments than to effect major changes to existing rules.

There are also important behavioral aspects to the issue. The way people process information, solve problems, and learn may be important for understanding institutional inertia and change. For example, people may systematically misperceive opportunities in a complex and changing environment or may be unaware of potentially beneficial institutional changes until the new institutions are "invented." North (2005) presents a framework in which economic actors have "mental models" which reflect their understanding of the world and which they use to evaluate the desirability of particular rule changes. Over time, as they learn about the world, they revise their mental models and may alter their perceptions of the effects of alternative rules and of the set of possible alternatives, providing an impetus for institutional change. This suggests that a key to understanding institutional change is an understanding of how people learn and revise their "mental models." Ongoing research in cognitive psychology and behavioral economics offers the promise of deepening our understanding of many aspects of institutional change.

Although viewing institutional change as the outcome of a deliberate, collective-choice process of rule creation yields many insights, it also leaves several important questions unanswered. In particular, it has difficulty explaining why, in many cases, formal rules are ignored or fail to produce their intended outcome. A key reason for this difficulty is the prevalence of "informal" constraints that are not a product of deliberate design and often vaguely defined. To clarify, let us distinguish three types of "informal rules."

First, the term "informal" is sometimes simply used to indicate that the rules are not written down or are not enforced by the state. A related interpretation emphasizes the importance of actors with specialized roles in enforcing formal rules (Milgrom et al. 1990).

Second, informal constraints are sometimes viewed as ethical codes or moral "norms" which are internalized and directly reflected in players' preferences.

Third, an important category of "informal rules" includes conventions – viewed as self-enforcing solutions to multiplayer coordination games (Sugden 1989) – and "social norms" which use a multilateral reputation mechanism and a credible threat of punishment to generate trust among members of a community (see, e.g., Kandori 1992; Greif 2006). Of course, these constraints may eventually come to be followed without rational evaluation and socially experienced as moral, ideological, or "cultural" rather than purely strategic constraints. However, if all others are following such rules, then even fully rational strategic players may also be induced to follow them. This is crucial because, even if most people follow the norm without rational evaluation, it is unlikely that a norm could evolve or survive if a rational mutant could achieve a higher payoff by deviating from it.

The second and third categories of informal rules (moral norms, conventions, and social norms) do, sometimes, change over time, but they do not fit easily into the collective-choice models because they generally evolve in a decentralized, "spontaneous" manner, rather than being deliberately designed and agreed to. This may be a serious shortcoming in some contexts because the evolution of informal rules is frequently an important part of the story of institutional change. Internalized moral codes, for example, may to some extent evolve to be compatible with prevailing formal rules, but they can also impact how formal rules change. For example, certain proposed rules may not be adopted because they are perceived as "unfair." This is another channel through which "history matters." Similarly, it is not uncommon for informal "rules" to originate as voluntary patterns of behavior that develop within a community and are later "formalized." Many aspects of commercial law, for example, derived from the codification of

merchant's practices which had evolved spontaneously (Milgrom et al. 1990; Kingston 2007).

Evolutionary Theories of Institutional Change

A large body of literature treats institutional change as an evolutionary process, in which new institutional forms periodically emerge and undergo some kind of decentralized selection process as they compete against alternative institutions. In the evolutionary theories, new rules or behaviors (mutations) may emerge from deliberate human actions (including learning, imitation, and experimentation), or they may develop "spontaneously" from the uncoordinated choices of many individuals. The key difference between evolutionary theories and the collective-choice theories discussed in the previous section has to do with the decentralized selection process which determines which rules ultimately become widely adopted. Those institutions that prove successful spread, by imitation or replication, while unsuccessful institutions die out. As a result, overall institutional change occurs "spontaneously," through the uncoordinated choices of many agents, rather than via a single, collective-choice or political mechanism.

Sometimes, an evolutionary model of institutional change is implicit. Williamson (2000)'s "Transactions cost economics," for example, *assumes* that the sets of rules ("governance structures") that can most efficiently govern any particular transaction (those that "minimize transactions costs") are those that will be observed. Implicitly, this rests on an assumption that competitive pressure would weed out inefficient forms of organization, as originally suggested by Alchian (1950), because more efficient organizational forms will yield higher profits and will therefore survive and be imitated. So, for example, if a change in production technology renders existing institutions inefficient, competitive pressures ensure that a new configuration of optimal institutions will emerge. This approach is an example of "functionalist" reasoning: to explain the attributes of an institution, we need to only ask what function it serves.

Although a functionalist approach can successfully explain many aspects of observed institutions, it does best in situations in which competition among institutional forms can plausibly operate to weed out inefficient rules and leads to a unique, optimal equilibrium. It has difficulty explaining why countries with similar technologies may use different institutions to govern apparently similar transactions, why inefficient institutions often seem to persist, or why less successful societies often fail to adopt the institutional structure of more successful ones.

The essence of these difficulties is that evolutionary processes frequently exhibit multiple equilibria. For example, credit cards are widely used by consumers because many merchants accept them for payment; and they are widely accepted because they are widely used. The resulting equilibrium is associated with a set of formal rules (credit card agreements), norms, and behaviors (carrying little cash), but we can easily envisage many other alternative institutional configurations, any of which, once established, would generate stable expectations and behavior within the context of a different set of "rules." That, is there may be multiple possible sets of self-enforcing rules ("conventions"), and there is no guarantee that the most efficient will be observed (Sugden 1989).

The possibility of multiple evolutionarily stable equilibria has two important, closely related consequences. First, observed institutions are not necessarily "efficient." And second, institutional change may exhibit "path dependence," in the sense that initial conditions and historical events can have a lasting impact on the institutions which are ultimately observed. For example, an institutional structure which was previously optimal might become sub-optimal as circumstances change, but without a coordinating device, such as legislation or the appearance of a "political entrepreneur," to engineer a change in the rules, the economy might remain stuck in the (now) sub-optimal equilibrium. Young (1996) uses an evolutionary framework to argue that historical accidents could lead to the selection of particular

conventions and argues that in the long run, the pattern of institutional change will follow a "punctuated equilibrium" process in which rapid switches between conventions are interspersed with long periods of stability.

Brousseau and Raynaud (2011) argue that many institutional arrangements begin as "private" local experiments, in which participation is voluntary, but that over time, through competition for adherents, economies of scale, and network effects, some (not necessarily optimal) institutions spread and emerge as "winners" and become "solidified" and formalized as part of the institutional environment, while participation in them becomes increasingly widespread and mandatory. Thus, they argue, some kinds of informal institutions can gradually "climb the ladder" and metamorphose into formal and permanent rules.

Blending Evolution and Design

Both the evolutionary and design-based approaches to understanding institutional change are useful in particular settings, but both are incomplete. Theories which view institutional change as the outcome of a centralized collective-choice process have difficulty explaining changes in informal constraints (such as social norms) that evolve in a decentralized manner, while evolutionary theories tend to neglect the role of collective action and the political process. This is not meant as a criticism of these theories: they have been developed to study a variety of different situations, and the assumptions and conclusions naturally reflect these differences. But in many real-world settings, both evolutionary processes and intentional design are at work, and it will often be difficult to cleanly separate the two. For example, gradual underlying changes in parameters, beliefs, or knowledge, which result from the spontaneous evolution of existing institutions over time, may give rise to deliberate attempts to design and implement new institutions; and following such attempts, competition or other evolutionary processes may subsequently play a role in determining whether particular institutional innovations survive and spread.

The question therefore naturally arises as to how to integrate these theories. To a large extent, this turns on the interaction between formal rules, which are generally deliberately designed (although there may also be evolutionary processes underlying their creation, as in the case of the common law), and informal rules, which are "much more impervious to deliberate policies" (North 1990: 6) and therefore (usually) evolve spontaneously.

Here, again, it is useful to begin with North (1990)'s seminal contribution. In North's account, formal rules change through a political process as a result of deliberate (though boundedly rational) actions by organizations and individual entrepreneurs, while informal rules evolve alongside, and as extensions of, formal rules. Informal rules play a key role in institutional change because they change slowly and cannot be changed deliberately. Following a change of formal rules, therefore, the informal rules which "had gradually evolved as extensions of previous formal rules" (ibid.: 91) survive the change, so that the result "tends to be a restructuring of the overall constraints – in both directions – to produce a new equilibrium that is far less revolutionary" (ibid.: 91). Essentially then, formal rules occupy the driving role in institutional change; informal constraints apply the brakes.

In general, there are multiple equilibria and no guarantee of an efficient outcome (North 1990: 80–81: 136). The process of institutional change is also path-dependent because individuals learn, organizations develop, and ideologies form in the context of a particular set of formal and informal rules (1990, chapter 9). These organizations then may attempt to change the formal rules to their benefit, and over time this in turn may (indirectly) affect the informal rules.

Roland (2004) distinguishes between "fast-moving" (political) institutions (akin to formal rules), which can be changed quickly and deliberately via the centralized political process, and "slow-moving" (cultural) institutions (akin to informal rules), which change slowly because change is continuous, evolutionary, and decentralized. He outlines his view of institutional change by analogy: tectonic pressures along fault

lines (changes in slow-moving institutions) build up continuously but slowly and then suddenly provoke an "earthquake" that causes abrupt and substantial changes in fast-moving institutions (i.e., formal rules). Thus, Roland's theory is, in a sense, the inverse of North's, in that changes in informal rules, rather than formal rules, are the main drivers of institutional change.

The "Equilibrium View" of Institutions

A growing body of recent research shifts the focus by identifying institutions with equilibrium patterns of behavior rather than the "rules" that induce the behavior (Greif and Kingston 2011). This "equilibrium perspective" emphasizes that the constraints that motivate individual behavior are ultimately derived from expectations about the behavior of other actors in various contingencies. In maximizing their welfare subject to these constraints, agents choose strategies which, in the aggregate, and perhaps unintentionally, give rise to expectations which reinforce the constraints on everyone else. Of course, these expectations may be summarized in the form of formal and informal "rules." But the constraints that motivate behavior – the "true" rules of the game – emerge as endogenous equilibrium outcomes, reflecting a socially constructed reality.

From the equilibrium perspective, institutional change becomes fundamentally not about changing rules, but about changing expectations; the essential goal of introducing new "rules" is to help players coordinate on one of the many possible equilibria by coordinating their beliefs about each other's expected behavior both on and off the path of play. In equilibrium, however, it is ultimately these behavioral expectations, rather than the prescriptive content of the rules themselves, that motivate compliance. And, for a variety of reasons, it is possible that an attempt to introduce new "rules" may fail to change these expectations and equilibrium patterns of behavior or may change them in unanticipated ways. The process of institutional change is consequential precisely because there are typically multiple equilibria.

A rule "forbidding" some behavior, for example, will be effective only if people generally expect others (including those charged with enforcing the rule) to act in a way which makes it effective.

A formal rule making one player a judge, president, or police officer does not change that player's set of physically feasible actions, but it *may* systematically alter people's perceptions about how those actions are to be interpreted and how other players will respond: *if* the rule is effective, then by virtue of her role, the judge can take actions and give orders which would not be followed if she were an ordinary citizen. In order for the rules to be effective, the behavior specified by the "rules" – including that of the "enforcers" of the rules – must correspond to an equilibrium in the underlying "game of nature."

Theories that define institutions as rules tend to obscure these possibilities because they consider the enforcement of rules separately from their content; they cannot explain why some rules are followed and others are not. For example, the legal speed limit on highways in Massachusetts is 65 miles per hour, but this limit is widely ignored. This is not to say that there are no "rules," however. Most cars travel around 70 mph and are (almost) never pulled over at this speed, whereas cars traveling at 80 mph sometimes are (and those traveling at 90 mph frequently are). What accounts for the difference between the behavior specified by the "formal rule" and the behavior actually observed? Asserting that there is an "informal rule" specifying the observed behavior is unsatisfactory; it merely assumes away what we would most like to explain. The equilibrium perspective offers a more satisfactory explanation: drivers' and police officers' behaviors constitute an overall equilibrium based on a broadly shared set of behavioral expectations. By making the "enforcement" of rules endogenous to the analysis, this approach enables the treatment of "formal" and "informal" constraints in a unified manner.

Both deliberate, centralized and evolutionary, decentralized institutional changes are compatible with the equilibrium view. Exogenous parameter shifts such as changes in technology or

preferences can disrupt an equilibrium, leading individuals and organizations to try to change the "formal rules" in order to achieve a coordinated shift of many players' beliefs about each others' strategies. Previous institutions also provide focal points which can affect equilibrium selection in novel situations (Sugden 1989). Alternatively, gradual changes in parameters might cause gradual adjustments to expectations and behavior. Since the formal rules remain unchanged, this kind of institutional change could, of course, be interpreted as changing "informal rules," but that merely labels the phenomenon without explaining it. Fundamentally, what is changing is a pattern of equilibrium behavior.

Greif and Laitin (2004) refer to parameters which are exogenous in the short run but which gradually change as a result of the play of the game as "quasi-parameters." Changes in quasi-parameters may either broaden the range of situations in which the existing pattern of behavior (institution) is an equilibrium or may undermine the existing institution, leading to an "institutional disequilibrium" and an impetus for institutional change. Institutional change, in this view, is highly path-dependent; as Greif and Laitin emphasize, knowledge, resource ownership, wealth distribution, and other "quasi-parameters" can all be affected by past institutions and affect both the future institutional choice set (the set of feasible equilibria) and the choice of institutions within that set.

Conclusion

The appropriate model for studying institutional change is largely a matter of context. For situations in which competition will tend to weed out inefficient institutional forms, functionalist explanations such as the "transaction cost" model are likely to be useful for explaining observed institutions. In situations in which changes in formal rules occur within a stable political context, and have relatively predictable effects on behavior, treating institutional change as an outcome of collective action and political maneuvering may be more suitable. However, this approach cannot explain why some formal rules become effective and others do not and tend to neglect the role of informal rules. The equilibrium view of institutions provides a more complete theory by treating both informal and formal rules, and their enforcement, within an integrated framework, and is therefore useful as a broad conceptual framework for understanding institutional change. However, it may introduce unnecessary complexity in the many real-world cases in which formal rules are relatively straightforward and effectively enforced.

Cross-References

- ▶ Constructivism, Cultural Evolution, and Spontaneous Order
- ▶ De Jure/De Facto Institutions
- ▶ Institutional Economics
- ▶ Path-Dependent Rule Evolution
- ▶ Political Economy
- ▶ Transaction Costs

References

Acemoglu D, Robinson JA (2005) The rise of Europe: Atlantic trade, institutional change, and economic growth. Am Econ Rev 95(3):546–579

Acemoglu D, Robinson JA (2006) Economic origins of dictatorship and democracy. Cambridge University Press, Cambridge

Alchian A (1950) Uncertainty, evolution and economic theory. J Polit Econ 58(3):211–221

Brousseau E, Raynaud E (2011) Climbing the hierarchical ladders of rules: a life-cycle theory of institutional evolution. J Econ Behav Organ 79(1):65–79

David PA (1994) Why are institutions the 'carriers of history'. Struct Chang Econ Dyn 5(2):205–220

David PA (2007) Path dependence, its critics and the quest for 'historical economics. In: Hodgson GM (ed) The evolution of economic institutions: a critical reader. Edward Elgar Press, Cheltenham, pp 120–142

Greif A (2006) Institutions and the path to the modern economy. Cambridge University Press, Cambridge

Greif A, Laitin D (2004) A theory of endogenous institutional change. Am Polit Sci Rev 98(4):633–652

Kandori M (1992) Social norms and community enforcement. Rev Econ Stud 59:63–80

Kantor SE (1998) Politics and property rights: the closing of the open range in the postbellum south. University of Chicago press, Chicago

Kingston C (2007) Marine insurance in Britain and America, 1720–1844: a comparative institutional analysis. J Econ Hist 67(2):379–409
Milgrom P, North D, Weingast B (1990) The role of institutions in the revival of trade: the law merchant, private judges, and the champagne fairs. Econ Polit 1:1–23
North D (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge
North D (2005) Understanding the process of economic change. Princeton University Press, Princeton
Ostrom E (2005) Understanding institutional diversity. Princeton University Press, Princeton
Roland G (2004) Understanding institutional change: fast-moving and slow-moving institutions. Stud Comp Int Dev 38(4):109–131
Sugden R (1989) Spontaneous order. J Econ Perspect 3(4):85–97
Williamson O (2000) The new institutional economics: taking stock, looking ahead. J Econ Lit 38:595–613
Young HP (1996) The economics of convention. J Econ Perspect 10(2):105–122

Further Reading

Aoki M (2001) Towards a comparative institutional analysis. MIT Press, Cambridge
Kingston C, Caballero G (2009) "Comparing theories of institutional change". J Inst Econ 5(2):151
Greif A, Kingston C (2011) Institutions: rules or equilibria? In: Caballero G, Schofield N (eds) Political economy of institutions, democracy and voting. Springer, Berlin
North D (1990) Institutions, institutional change and economic performance. Cambridge University Press, Cambridge

Institutional Complementarity

Fabio Landini[1] and Ugo Pagano[2,3]
[1]LUISS University, Rome, Italy
[2]University of Siena, Siena, Italy
[3]Central European University, Budapest, Hungary

Abstract

Institutional complementarity refers to situation of interdependence among institutions. The present article presents a formal representation of institutional complementarity and discusses several economic applications of this concept.

Definition

Even if several definitions have been proposed in the literature (Crouch et al. 2005) they share the idea that institutional complementarity refers to situations of interdependence among institutions. Institutional complementarity is frequently used to explain the degree of institutional diversity that can be observed across and within socioeconomic system and its consequences on economic performance. In particular, the concept of institutional complementarity has been used to illustrate why institutions are resistant to change and why introducing new institutions into a system often leads to unintended, sometimes suboptimal, consequences.

Institutional Complementarity

The canonical formal representation of the concept of institutional complementarity is due to Aoki (2001) and relies on the theory of supermodular games developed by Milgrom and Roberts (1990). The basic structure of the model takes the following form.

Let us consider a setting with two institutional domains, A and B, and two sets of agents, C and D that do not directly interact with each other. Nevertheless, an institution implemented in one domain parametrically affects the consequences of the actions taken in the other domain. For instance, A can be associated with the type of ownership structure prevailing in a given country and B with the structure of labor rights. For simplicity, we assume that the technological and natural environment is constant.

Suppose that the agents in domain A can choose a rule from two alternative options: A^1 and A^2; similarly, agents in domain B can choose a rule from either B^1 or B^2. For simplicity, let us assume that all agents in each domain have an identical payoff function $u_i = u(i \in C)$ or $v_j = v(j \in D)$ defined on binary choice sets of their own, either $\{A^1; A^2\}$ or $\{B^1; B^2\}$, with another sets as the set of parameters. We say that an (endogenous) "rule" is institutionalized in a

domain when it is implemented as an equilibrium choice of agents in the relevant domains.

Suppose that the following conditions hold:

$$u(A^1;B^1) - u(A^2;B^1) \geq u(A^1;B^2) - u(A^2;B^2) \quad (1)$$

$$v(B^2;A^2) - v(B^1;A^2) \geq v(B^2;A^1) - v(B^1;A^1) \quad (2)$$

for all i and j. The latter are the so-called supermodular (complementarity) conditions. Equation 1 implies that the "incremental" benefit for the agents in A from choosing A^1 rather than A^2 increases as their institutional environment in B is B^1 rather than B^2. Equation 2 implies that the "incremental" benefit for agents in B from choosing B^2 rather than B^1 increases if their institutional environment in A is A^2 rather than A^1. Note that these conditions are concerned with the property of incremental payoffs with respect to a change in a parameter value. They do not exclude the possibility that the level of payoff of one rule is strictly higher than that of the other for the agents of one or both domain(s) regardless of the choice of rule in the other domain. In such a case the preferred rule(s) will be implemented autonomously in the relevant domain, while the agents in the other domain will choose the rule that maximizes their payoffs in response to their institutional environment. Then the equilibrium of the system comprised of A and B – and thus the institutional arrangement across them – is uniquely determined by preference (technology).

However, there can also be cases in which neither rule dominates the other in either domain in the sense described above. If so, the agents in both domains need to take into account which rule is institutionalized in the other domain. Under the supermodularity condition, there can then be two pure Nash equilibria (institutional arrangements) for the system comprised of A and B, namely, $(A^1; B^1)$ and $(A^2; B^2)$. When such multiple equilibria exist, we say that A^1 and B^1, as well as A^2 and B^2, are "institutional complements."

If institutional complementarity exists, each institutional arrangement characterizes as a self-sustaining equilibrium where no agent has an inventive to deviate. In terms of welfare, it may be the case that possible overall institutional arrangements are not mutually Pareto comparable or that one of them could be even Pareto suboptimal to the other. In these cases, history is the main force determining which type of institutional arrangements is likely to emerge, with the consequence that suboptimal outcomes are possible.

Suppose, for instance, that $(A^2; B^2)$ is a Pareto-superior institutional arrangement in which $u(A^2; B^2) > u(A^1; B^1)$ and $v(B^2; A^2) > v(B^1; A^1)$. However, for some historical reason A^1 is chosen in domain A and becomes an institutional environment for domain B. Faced with this institutional environment, agents in domain B will correspondingly react by choosing B^1. Thus the Pareto-suboptimal institutional arrangement $(A^1; B^1)$ will result. This is an instance of coordination failure in the presence of indivisibility.

Obviously, there can also be cases where $u(A^2; B^2) > u(A^1; B^1)$ but $v(B^1; A^1) > v(B^2; A^2)$. This is an instance where the two viable institutional arrangements cannot be Pareto ranked. Agents exhibit conflicting interests in the two equilibria, and the emergence of one institutional arrangement as opposed to the other may depend on the distribution of decisional power. If for some reasons agents choosing in domain A have the power to select and enforce their preferred rule, arrangement $(A^2; B^2)$ is the most likely outcome. Alternatively, agents choosing in domain B will force the society to adopt $(B^1; A^1)$.

Pagano (1992) and Pagano and Rowthorn (1994) are two of the earliest analytical contributions to institutional complementarity. In their models, the technological choices take as parameters property rights arrangements whereas the latter are made on the basis of given technologies. The complementarities of technologies and property rights create two different sets of organizational equilibria. For instance, strong rights of the capital owners and a technology with a high intensity of specific and difficult to monitor capital are likely be institutional complements and define one possible organizational equilibrium. However,

also strong workers' rights and a technology characterized by a high intensity of highly specific labor can be institutional complements and define an alternative organizational equilibrium. The organizational equilibria approach integrates the approaches *à la* Williamson (1985), which have pointed out the influence of technology on rights and safeguards, and the views of the radical economists (see, for instance, Braverman 1974), who have stressed the opposite direction of causation. The complementarities existing in the different organizational equilibria integrate both directions of causation in a single analytical framework. A similar approach has been used to explain organizational diversity in knowledge-intensive industries, such as software (Landini 2012, 2013).

Institutional complementarities characterize also the relations between intellectual property and human capital investments. Firms owning much intellectual property enjoy a protection for their investments in human capital, which in turn favor the acquisition of additional intellectual property. By contrast other firms may find themselves in a vicious circle where the lack of intellectual property inhibits the incentive to invest in human capital and low levels of human capital investments involve that little or no intellectual property is ever acquired (Pagano and Rossi 2004).

Less formal approaches to institutional complementarities have also been adopted. In their seminal contribution, Hall and Soskice (2001) develop a broad theoretical framework to study the institutional complementarities that characterize different varieties of capitalism. Having a specific focus on the institutions of the political economy, the authors develop an actor-centered approach for understanding the institutional similarities and differences among the developed economies. The varieties of capitalism approach has inspired a large number of applications to the political economy field. To give some examples, Franzese (2001) and Höpner (2005) investigate the implications for industrial relations; Estevez-Abe et al. (2001) use the approach to analyze social protection; Fioretos (2001) considers political relationships, international negotiations, and national interests; Hall and Gingerich (2009) study the relationship among labor relations, corporate governance, and rates of growth; Amable (2000) analyzes the implications of institutional complementarity for social systems of innovation and production.

In addition to institutional variety, the notion of institutional complementarity has also motivated studies on institutional change (Hall and Thelen 2009; Boyer 2005). In these works institutional complementarity is often presented as a conservative factor, which increases the stability of the institutional equilibrium (Pagano 2011). In the presence of institutional complementarity change requires the simultaneous variation of different institutional domains, which in turn demands high coordination among the actors involved. Sometime, institutions themselves can act as resources for new courses of action that (incrementally) favor change (Deeg and Jackson 2007).

Alongside contributions on the distinct models of capitalism, the concept institutional complementarity has found application also in other domain of analysis. Aoki (1994), for instance, studies the role of institutional complementarity in contingent governance models of teams. Siems and Deakin (2010) rely on an institutional complementarity approach to investigate differences in the business laws governing in various countries. Gagliardi (2009) argue in favor of an institutional complementarity relationship between local banking institutions and cooperative firms in Italy. Bonaccorsi and Thuma (2007), finally, use the idea of institutional complementarity to investigate inventive performance in nano science and technology.

References

Amable B (2000) Institutional complementarity and diversity of social systems of innovation and production. Rev Int Polit Econ 7(4):645–687

Aoki M (1994) The contingent governance of teams: analysis of institutional complementarity. Int Econ Rev 35(3):657–676

Aoki M (2001) Toward a comparative institutional analysis. MIT Press, Cambridge

Bonaccorsi A, Thoma G (2007) Institutional complementarity and inventive performance in nano science and technology. Res Policy 36(6):813–831
Boyer R (2005) Coherence, diversity, and the evolution of capitalisms: the institutional complementarity hypothesis. Evol Inst Econ Rev 2(1):43–80
Braverman H (1974) Labor and monopoly capital: the degradation of work in the twentieth century. Monthly Review Press, New York
Crouch C, Streek W, Boyer R, Amable B, Hall P, Jackson G (2005) Dialogue on "Institutional complementarity and political economy". Soc Econ Rev 3:359–382
Deeg R, Jackson G (2007) Towards a more dynamic theory of capitalist variety. Soc Econ Rev 5(1): 149–179
Estevez-Abe M, Iversen T, Soskice D (2001) Social protection and the formation of skills: a reinterpretation of the welfare state. In: Hall PA, Soskice D (eds) Varieties of capitalism: the institutional foundations of comparative advantage. Oxford University Press, New York, pp 145–183
Fioretos O (2001) The domestic sources of multilateral preferences: varieties of capitalism in the European Community. In: Hall PA, Soskice D (eds) Varieties of capitalism: the institutional foundations of comparative advantage. Oxford University Press, New York, pp 213–244
Franzese RJ (2001) Institutional and sectoral interactions in monetary policy and wage/price-bargaining. In: Hall PA, Soskice D (eds) Varieties of capitalism: the institutional foundations of comparative advantage. Oxford University Press, New York, pp 104–144
Gagliardi F (2009) Financial development and the growth of cooperative firms. Small Bus Econ 32(4): 439–464
Hall PA, Gingerich DW (2009) Varieties of capitalism and institutional complementarities in the political economy: an empirical analysis. Br J Polit Sci 39(3):449–482
Hall PA, Soskice D (2001) Varieties of capitalism: the institutional foundations of comparative advantage. Oxford University Press, New York
Hall PA, Thelen K (2009) Institutional change in varieties of capitalism. Soc Econ Rev 7(1):7–34
Höpner M (2005) What connects industrial relations and corporate governance? Explaining institutional complementarity. Soc Econ Rev 3(1):331–358
Landini F (2012) Technology, property rights and organizational diversity in the software industry. Struct Chang Econ Dyn 23(2):137–150
Landini F (2013) Institutional change and information production. J Inst Econ 9(3):257–284
Milgrom P, Roberts J (1990) Rationalizability, learning and equilibrium in games with strategic complementarities. Econometrica 58(6):1255–1277
Pagano U (1992) Organizational equilibria and production efficiency. Metroeconomica 43:227–246
Pagano U (2011) Interlocking complementarities and institutional change. J Inst Econ 7(3):373–392
Pagano U, Rossi MA (2004) Incomplete contracts, intellectual property and institutional complementarities. Eur J Law Econ 18(1):55–67
Pagano U, Rowthorn R (1994) Ownership, technology and institutional stability. Struct Chang Econ Dyn 5(2):221–242
Siems M, Deakin S (2010) Comparative law and finance: past, present, and future research. J Inst Theor Econ 166(1):120–140
Williamson OE (1985) The economic institutions of capitalism. The Free Press, New York

Institutional Economics

Ringa Raudla
Faculty of Social Sciences, Ragnar Nurkse School of Innovation and Governance, Tallinn University of Technology, Tallinn, Estonia

Abstract

Institutional economics is interested in the interactions between institutions and the economy: how institutions influence the functioning, performance, and development of the economy and, in turn, how changes in the economy influence the institutions. Institutional economics studies the impact of institutions on economy, how institutions evolve, and how they could be improved. In furthering institutional economics, more extensive exchanges between the communities of institutional economics and law and economics would be fruitful. Those interested in institutional economics should certainly be more aware of developments in law and economics and utilize these insights in their further research – and vice versa.

Definition

Institutional economics is interested in the interactions between institutions and the economy: how institutions influence the functioning, performance, and development of the economy and, in turn, how changes in the economy influence the institutions. Institutional economics

studies the impact of institutions on economy, how institutions evolve, and how they could be improved.

Introduction

For readers from outside the discipline of economics, the adjective "institutional" in front of "economics" in the title of this entry may appear somewhat redundant. It would seem obvious that in order to understand what is going on in the economy, it is necessary to examine the institutions that the economy is embedded in. Thus, one could ask: why it is necessary to emphasize the term "institutions" in "institutional economics" given that "economics" should, logically, already include the analysis of "institutions"?

The reason for including the term "institutional" in front of "economics" is that neoclassical economics, which has constituted the economic mainstream since the mid-twentieth century, has paid only very limited attention to institutions. This has, at least partly, been due to the fact that economics came to be defined by its *method* (i.e., formalistic analysis of rational choice), rather than its *object* of analysis (i.e., the economy).

Fortunately, despite the lure of the mainstream economics in which the method was given prominence over the subject matter, a sufficient number of economists have been interested in real-life economies. In studying the *economy* as a subject matter, however, institutions cannot be ignored for too long before the analysis becomes stalled. Thus, in recent decades, institutions have returned to economic analysis, with an increasing number of economists paying attention to the role of institutions when explaining what is going on in the economy.

The law and economics movement has certainly played an important role in bringing "institutions" (like law) back into economics. Indeed, institutional economics and law and economics are closely related: they have similar intellectual roots, common pioneers, and overlapping research agenda. Law is, obviously, one of the most important "institutions" that are studied in institutional economics. Some would even say that "law and economics" could be viewed as one of the subfields or "building blocks" of institutional economics.

What Is Institutional Economics?

Most broadly speaking, institutional economics is interested in the interactions between institutions and the economy: how institutions influence the functioning, performance, and development of the economy and, in turn, how changes in the economy influence institutions. Institutional economics studies the impact of institutions on the economy, how institutions evolve, and how they could be improved.

When studying the economic system, institutional economics assumes that "institutions matter" since institutions are the key element of any economy and should hence be placed at the center of analysis. Various definitions of the term "institution" have been offered by different authors. Broadly speaking, institutions can be defined as formal and informal rules, including their enforcement mechanisms. Douglass North, who can be considered as one of the "bridge-builders" between the old and new institutional economics, has defined institutions as "the rules of the game in a society" (1990, p. 3) or, more specifically, "humanly devised constraints that structure political, economic and social interaction" (1991, p. 97). According to North, institutions consist of both "informal constraints" (e.g., customs, traditions, and codes of conduct) and "formal rules" (e.g., constitutions and laws).

Some analysts (especially in the new institutionalist camp) have considered it useful to distinguish between different levels of analysis in examining institutions (see, e.g., Williamson 2000; North 1990): (1) the "highest" or so-called "social embeddedness" level entails norms, customs, and traditions; (2) the next level – the "institutional environment" – comprises of constitutions, laws, and political institutions; and (3) the lowest level of analysis looks at "institutional arrangements" or "governance arrangements" (e.g., vertical integration, franchising) and "organizations" (e.g., political parties, firms,

and trade unions). Institutional economics considers all these different levels of analysis as worthy pursuits (including interactions between the different levels), though admittedly, a lot more progress has been made on the third level of analysis than on the first two, where a lot of research still remains to be done.

Roots of Institutional Economics

The roots of institutional economics go back (at least) to the German historical school (represented, e.g., by Gustav von Schmoller, Wilhelm Roscher, Werner Sombart, and Max Weber) and the American institutionalist school (represented, e.g., by John Commons, Thorstein Veblen, and Wesley Mitchell) (for more detailed discussions, see, e.g., Medema et al. 1999; Rutherford 1994). The historical school dominated the discipline of economics in Germany from 1840s to 1940s, whereas the American institutionalist school had its peak during the interwar period. While there were important differences between these two schools and also between the individual thinkers within them, what they had in common was the focus on the role of institutions in shaping economic outcomes and their critique of the neoclassical approach to economic analysis (especially the formalistic aspects of it). The German historical school underscored the importance of sensitivity to specific cultural and historical circumstances in economic analysis (resulting in their emphasis on using empirical data to ground economic theories). The American institutionalist school emphasized the relevance of institutions (including the role of the laws and the state) in analyzing the economy. Commons (1924), for example, examined the legal underpinnings of the capitalist economic system; he demonstrated how evolutionary changes in the economic domain (e.g., with regard to what types of activities were deemed reasonable) facilitated specific changes in laws (e.g., the transformation of how the concept of property was legally defined) and how these changes, in turn, facilitated specific forms of economic activity. The American institutionalist school was also skeptical of the notion of the fixed preferences of individuals and emphasized that individual preferences can be shaped by the institutions that surround them (e.g., via forming habits, as argued by Thorstein Veblen). In addition, the institutionalists criticized the static approach of the neoclassical economics and emphasized the evolutionary nature of economy (and hence the focus on change, technology, and innovation in economic analysis) (see, e.g., Veblen 1898).

In the postwar period, the popularity of the institutionalist approaches waned and economics became dominated by neoclassical economics. However, the institutionalist traditions were carried on by economists like Clarence Ayres, Karl Polanyi, John Kenneth Galbraith, Allan Gruchy, Simon Kuznets, Gunnar Myrdal, Ragnar Nurkse, Joseph Schumpeter, and others. In law and economics, the institutionalist traditions have been carried on by Allan Schmid, Warren Samuels, Nicolas Mercuro, and Steven Medema.

Different Approaches Within the "Modern" Institutional Economics

Until a decade or so ago, authors writing about "institutional economics" considered it necessary to distinguish between "old" and "new" institutional economics (for a more detailed discussion of the differences between these two camps, see, e.g., Rutherford 1994). "Old" institutionalism referred to researchers (e.g., Wendell Gordon, Allan Gruchy, Philip Klein, Marc Tool, Warren Samuels, Allan Schmid) following the traditions of John Commons, Thorstein Veblen, Wesley Mitchell, and others. The "new" institutional economics referred to the developments in economics that started in 1960s (led by Ronald Coase, Douglass North, and Oliver Williamson) when institutions were (at least to some extent) "brought back in" to the economic mainstream. Many authors in the emerging new institutional economics camp (e.g., Robert Sugden, Andrew Schotter, Mancur Olson, and Richard Posner) attempted to use the analytical tools of neoclassical theory to explain the emergence and impacts of institutions, with a specific attention to transaction costs, property rights, and contractual relations. In more recent years, however, the academics in the new institutionalist camp have increasingly moved away from the assumptions of neoclassical economics.

While one can still observe differences between the "old" and "new" camps, one can also talk more generally about (modern) institutional economics. In the light of the recent convergences between the camps (see, e.g., Dequech 2002), it would be helpful to offer a more *synthesized* view of what institutional economics is about. Thus, this entry tries to delineate the "common" ground of different institutional approaches by focusing on issues that many (if not most) researchers involved in doing research in institutional economics would agree are important.

Still, it is worth keeping in mind that institutional economics entails a rather diverse (and to some extent also conflicting) set of approaches. These different research streams vary with regard to the substantive questions studied and the methodology applied. Thus, institutional economics is not a single, unified, all-embracing, and well-integrated theory, proceeding from a set of common assumptions and hypotheses. Instead, it consists of different "building blocks," coming from different traditions. Furthermore, it is worth emphasizing that institutional economics is an openly interdisciplinary endeavor, which draws on other disciplines (like sociology, psychology, anthropology, history, political science, public administration, and law) in order to understand and explain the role of institutions in economic life.

Differences Between Institutional Economics and Neoclassical Economics

Although at least some of the topics that used to be the playground of institutional economics have gradually found their way into the economic mainstream and there is a growing consensus about the importance of institutions in influencing economic growth (see, e.g., Acemoglu et al. 2005), it is still too early to say that "we are all institutionalists now." Hence, a few remarks on how the *institutional* approach in economics differs from the *neoclassical* approach are still necessary. It is worth keeping in mind, though, that the different institutionalist camps differ somewhat with regard to their "distance" from the orthodox neoclassical economics: those who are closer to the "old" institutionalist traditions are further removed from the neoclassical assumptions than those in the "new" institutionalist camp.

In sum, the differences between the (mainstream) neoclassical economics and institutional economics are the following (for a more systematic comparison, see, e.g., Hodgson 1988; Medema et al. 1999):

First, while neoclassical economics proceeds from the assumption of "rational" individuals who maximize their utility (*homo oeconomicus*), the institutionalist approach takes a more realistic view of individual behavior: it regards individuals as being purposeful but only *boundedly rational*. Unlike orthodox economics, institutional economics emphasizes the importance of severe information problems (including uncertainty about the future) and the costs involved in obtaining necessary information. Proponents of institutional economics have hence criticized the orthodox approaches for simply "assuming away" the information problems and "assuming" perfect knowledge.

Second, in contrast to the neoclassical approach of treating the tastes and preferences of individuals as "given" or "fixed" (at least for the purposes of analysis), most institutional economists view the individuals as "social beings" and hence consider it necessary to proceed from the assumption that preferences are *malleable* and that *changes* in preferences should be analyzed as well, including the role of institutions in molding individual preferences and purposes (via changes in habits, as argued by Hodgson 1988, 1998).

Third, while neoclassical economics is concerned with states of *equilibria* and equilibrium-oriented theorizing (focusing on "mechanistic" optimization under static constraints), most institutional economists prefer to take a more evolutionary view of economic phenomena and also institutions. They emphasize that economic development has an *evolutionary* nature and hence prefer dynamic modes of theorizing, with a focus on longer-run processes of continuity and change, entailing path dependencies, transformations, and learning over time. Institutional economics is also interested in the

evolutionary nature of the interactions between institutions and the economy. Further, while the neoclassical economics treats technology as "given" (or exogenous), institutional economics emphasizes the importance of examining the role of technological changes (and their interactions with institutions) in economic development.

Fourth, while neoclassical economics tends to treat the use of *power* as given (and accept the power structure as it is), at least some institutional economists (especially those with closer ties to the "old" institutional economics) are concerned with how power is actually deployed in the economic, political, and societal settings. Power is deemed important because power relations influence who gets to shape the institutions (including legal rules), whose values dominate, and whose "interests" are to be regarded as "rights." The allocation of rights, in turn, would influence the distribution of power in society (see, e.g., Acemoglu et al. 2005; Furubotn and Richter 1997; Medema et al. 1999).

Why and How Do Institutions Matter?

Most generally speaking, institutions "matter" for the economy because the structures entailed in the institutions influence the behavior of the economic actors, which in turn influences the functioning and performance of the economy. The influence from the "institutions" to the "economy," however, is not unidirectional: changes in the economy can bring about changes in institutions as well and, hence, it would be more accurate to talk about mutual interactions between institutions and the economy (see, e.g., Medema et al. 1999). In other words, we should not take a deterministic view of institutions according to which institutions always *determine* the actions of individuals: the causal arrow can go in the other direction as well. "Actors and structure, although distinct, are thus connected in a circle of mutual interaction and interdependence" (Hodgson 1998, p. 181).

How do institutions influence the behavior of economic actors? There are different ways to answer the question. The answer closest to mainstream economics is that institutions shape the "choice set" available to the economic actors and "structure the incentives" of the actors (hence making a certain course of action more attractive than other courses) and thus steer individual behavior via affecting the costs and benefits associated with different types of actions (including engaging in different types of economic activities) (see, e.g., Acemoglu et al. 2005; Eggertsson 1990; Furubotn and Richter 1997; North 1991). Other answers point to the more "sociological" and deeper "psychological" mechanisms and emphasize role of institutions in shaping the habits and, through that, also the preferences of individuals, which, in turn, would influence their choices and actions (see, e.g., Hodgson 1988, 1998).

At the most basic level – and this is something that all institutionalists agree with – institutions influence the interactions of economic actors by providing "order" and reducing uncertainty in exchange. Given that institutions outline the "rules of the game" (which provide boundaries, constraints, and patterns for behavior), they provide economic actors with information about the potential behavior of *other* actors and hence help to establish baseline conditions for interactions between economic agents. Hence, institutions allow individuals to make reliable predictions about what *other* economic agents are likely to do in any given circumstance, which allows them to proceed with decision-making, negotiations, and exchange with at least some level of certainty (see, e.g., Hodgson 1988; Kasper and Streit 1998; Nelson and Sampat 2001; North 1991). Thus, institutions can both *constrain* and *enable* individual actions, e.g., by providing information about the behavior of others, defining pathways for doing things, allowing coordinated actions, and limiting opportunistic and arbitrary behavior.

Many institutional economists have emphasized that institutions influence the size of *transactions costs* associated with exchange relations (see, e.g., Furubotn and Richter 1997; Kasper and Streit 1998; North 1991). Transaction costs involve different cost associated with exchange processes, including search and information costs (e.g., discovering what one wants to buy, who the sellers are, and what the prices are),

bargaining and decision costs (e.g., associated with drawing up a contract), and policing and enforcement costs (see, e.g., Coase 1960; Furubotn and Richter 1997; North 1991). For example, provisions of contract law can help to lower the costs of concluding and enforcing contracts (e.g., the possibility to turn to courts in case of a breach reduces the need undertake "private" safeguarding measures by the parties themselves) and to hence facilitate impersonal contracting between strangers.

It has to be emphasized, however, that the institutions that have evolved in any given country do not necessarily *guarantee* a well-functioning economy and fast economic development. The institutions that have emerged can also be highly inefficient and entail elements that clearly hinder technological advances and economic development and growth (Freeman and Perez 1988; North 1990, 1991; Nelson and Sampat 2001).

Some Examples of How Institutions Influence Economic Performance

While lot of research still needs to be undertaken in order to achieve fuller understanding of the role of institutions in economic development, a number of insightful studies have been conducted so far. A complete overview of these achievements is certainly beyond the scope of this entry. Thus, the examples below constitute only a small portion of the body of research in institutional economics and should be viewed as indicative rather than exhaustive.

Markets as Institutions

Mainstream economics tends to treat the market as some sort of a "natural" feature of a social domain, an aggregate of individual bargains, resulting from free exchange between economic agents – almost as "an ether in which individual and subjective preferences relate to each other, leading to the physical exchange of goods and services," independent of institutions (Hodgson 1988, p. 178). In contrast, for most institutional economists, the market should be conceived of as an institution (or a set of institutions), involving "social norms, customs, instituted exchange relations and – sometimes consciously organized – information networks" (Hodgson 1998, p. 181). Institutional economics emphasizes that market institutions (and the institutions the market is embedded in) can play an important role in lowering transaction costs and hence facilitate more exchange relations.

Institutionalist research has also examined the role of the state in creating (what mainstream economists call) "free markets." Karl Polanyi (1944), for example, shows, in his study of the Industrial Revolution in Great Britain, that the development of free markets in the eighteenth and nineteenth centuries actually involved a significant increase in the activities of the government: more legislation was called for and more administrators needed to monitor and safeguard the free working of the market system. In other words, the creation of "free markets" implied an increase in the control, regulations, and intervention by the state. The same applies today: "every successful market economy is overseen by a panoply of regulatory institutions, regulating conduct in goods, services, labor, assets, and financial markets" (Rodrik 2000, p. 7). The "freer" the markets, the greater the vigilance that may be required from the regulatory institutions (e.g., in the field of antitrust, financial regulation, securities legislation, etc.) (ibid).

In the light of these findings, some institutionalists emphasize that the dichotomy between regulation vs deregulation (or intervention vs nonintervention or "more" vs "less government") is false. Instead, one should ask which *type* of regulation and intervention the state is engaged in (and for what ends) and whose "interests" are protected as "rights" by the state (Hodgson 1988; Medema et al. 1999). For example, if the government relaxes regulations pertaining to workplace safety, it expands the set of rights of employers and narrows the set of rights of employees – and vice versa when these regulations are toughened. In either case, the government is "present" – via the legal framework and the mechanisms of enforcement (Medema et al. 1999).

Property Rights

An important set of institutions that has captured the attention of many institutional economists – both in the "old" and "new" camps, from Commons (1924) to North (1990) – involves *property rights*. Again, while neoclassical economics takes property rights as "given" (i.e., perfectly defined), institutional economists take a much closer look at the actual definition, delineation, allocation, and enforcement of property rights and how these influence exchange relations and other economic activities. Institutionalists from different traditions agree that the specific content of property rights (e.g., control rights over assets or resources) and the way they are enforced influence the allocation and use of resources. As Rodrik (2000, p. 6) puts it, "an entrepreneur would not have an incentive to accumulate and innovate unless s/he has adequate control over the return of the assets that are thereby produced or improved." Thus, for example, when property rights are not credibly secured (e.g., when there is a threat of expropriation by the government or unilateral seizure by another private actor), entrepreneurs are less likely to adjust (efficiently) to changes in technology, to invest (e.g., in research and development, which facilitates technological change), and to innovate. In contrast, secure property rights encourage firms to make higher value-added investments with longer-term time horizons (Keefer and Knack 1997).

Institutional economists have also examined the role of the *state* in protecting property rights. On the one hand, it is emphasized that protection of individual property entails legal structures for recognizing, adjudicating, and enforcing these rights, which can be provided by the state (e.g., Sened 1997). On the other hand, it is argued (especially by the new institutionalists) that the state can also pose a danger to private property through expropriation (Furubotn and Richter 1997). Bringing these two arguments together, Douglass North (e.g., 1990, 1991) has argued that economic development is fostered by an institutional environment in which the state is sufficiently strong to protect the private parties from seizing each others' property but can at the same time make a credible commitment not to expropriate the very same property it is defending and securing. As Hodgson (2004) puts it, "For private property to be relatively secure, a particular form of state had to emerge, countered by powerful and multiple interest groups in civil society. This meant a pluralistic state with some separation of powers, backed up by a plurality of group interests in the community at large." In his empirical studies, Douglass North has argued that the establishment of clear and secure property rights played a major role in the economic development and rise of the "West." Establishing secure property rights (with a credible commitment by the state to respect and secure them) allowed, for example, the emergence of capital markets and the employment of technology necessary for industrial production (see, e.g., North 1990, 1991). Many other studies (e.g., Acemoglu et al. 2005; Keefer and Knack 1997) have confirmed that finding.

Another discussion concerning property rights pertains to the question of whether the policy instrument of more extensive allocation of property rights implies "more" state or "less" state. Some economists in the new institutionalist camp (e.g., Demsetz, Alchian) regard clearer definition and allocation of property rights (especially the extension of private property) as "solutions" to different types of market failures, hence allowing the "lessening" of the need for government intervention. Hodgson (1988, p. 152), in contrast, has pointed out that by expanding the domain of formal property rights, the state still remains engaged, but in a different way – through the extension of litigational activity. Chang (2007), among others, has also warned us of the dangers of using the institutional prescription of "private property rights" as the main "solution" for guaranteeing economic development (see also Rodrik 2000). Indeed, a whole stream of research examines the conditions under which different types of ownership – private property, common property, state property, and various hybrid forms (like the township and village enterprises in China) – lead to optimal use of resources. Elinor Ostrom (e.g., 1990), for example, has shown that in the case of common-pool resources, common property can (when

combined with suitable institutional arrangements) lead to a better use of natural resources (e.g., fish stock, forests, water) than either privatization or nationalization.

Political Institutions and Bureaucracy

One of the building blocks in the institutionalist literature looks at the impacts of specific features of *political institutions* (including constitutions) on economic performance. The starting point for many of these studies is that the balance and separation of powers and the number and power of veto players are likely to influence the general character of policy action (including the levels of decisiveness and credibility) and also specific features of policies and laws the governments adopt, which, in turn, can influence economic performance. For example, it has been examined how the level of democracy (and the level of inclusiveness in governance) but also specific constitutional features – like government type (presidential vs parliamentary), electoral rules (e.g., plurality vs proportional), the organization of the judiciary, and vertical separation of powers – influence economic performance (see, e.g., Acemoglu et al. 2005; Persson and Tabellini 2003; Rodrik 2000).

Yet another stream examines the role of administrative structure, public administration, and the characteristics of bureaucracy in the economic growth and development. Evans and Rauch (1999) show that economic growth is higher in those developing countries where the bureaucracies entail more Weberian elements (e.g., recruitment based on merit and predictable long-term career paths). They argue (drawing, e.g., on Weber [1904–1911] 1968 and also Polanyi 1944) that merit-based recruitment and long-term careers facilitate higher competence of public administrators, lower levels of corruption, and long-term orientation. These factors, in turn, facilitate the design and implementation of policies that can help to promote growth, e.g., the provision of long-term (public) investments that complement those made in the private sector and helping private sector actors to overcome coordination and information problems (see also Rodrik 2000; Wade 1990).

Concluding Remarks

Despite an increasing number of studies examining the links between institutional setting and economic performance, we are only at the beginning of the journey to understand the interrelations involved and which institutions constitute a "good fit" in different countries and contexts. As Chang (2007, p. 3) puts it: "We are still some way away from knowing exactly which institutions in exactly which norms are necessary, or least useful, for economic development in which contexts."

As the experience of transition, emerging, and developing economies has demonstrated, in further studies it would be necessary to explore the effects of different *configurations* of (complementary) institutions rather than examining the effects of specific institutions (e.g., the establishment of private property rights or the adoption of a new contract law) in isolation.

Also, the relationships between informal and formal institutions still need to be examined further. It is clear that the effectiveness of "formal" institutions (e.g., laws and regulations) depends on whether they fit sufficiently well with the "informal" institutions (like norms and customs), which in turn influences, for example, how well institutional transplants (from one country to another or implementing the "best practice" blueprints) can work. However, we are still far from completely understanding how informal forms emerge and persist, how such informal norms interact with formal norms, and how that, in turn, influences specific economic activities in a given country.

Finally, we still have only limited knowledge of how institutions conducive to economic development in specific contexts can be "built." These are all important arguments for undertaking more in-depth qualitative and comparative studies in the field.

In furthering institutional economics, more extensive exchanges between the communities of institutional economics and law and economics would be fruitful. Those interested in institutional economics should certainly be more aware of developments in law and economics and utilize these insights in their further research – and vice versa.

References

Acemoglu D, Johnson S, Robinson JA (2005) Institutions as a fundamental cause of long-run growth. In: Aghion P, Durlauf SN (eds) Handbook of economic growth, vol 1A. Elsevier, Amsterdam, pp 385–472

Chang H-J (ed) (2007) Institutional change and economic development. United Nations University Press, Tokyo

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

Commons JR (1924) Legal foundations of capitalism. Macmillan, New York

Dequech D (2002) The demarcation between the "old" and the "new" institutional economics: recent complications. J Econ Issues 36(2):565–572

Eggertsson T (1990) Economic behavior and institutions: principles of neoinstitutional economics. Cambridge University Press, Cambridge

Evans P, Rauch JE (1999) Bureaucracy and growth: a cross-national analysis of the effects of "Weberian" state structures on economic growth. Am Sociol Rev 65(5):748–765

Freeman C, Perez C (1988) Structural crises of adjustment, business cycles, and investment behavior. In: Dosi G (ed) Technical change and economic theory. Pinter Press, London, pp 38–66

Furubotn EG, Richter R (1997) Institutions and economic theory: the contribution of the new institutional economics. The University of Michican Press, Ann Arbor

Hodgson GM (1988) Economics and institutions: a manifesto for a modern institutional economics. University of Pennsylvania Press, Philadelphia

Hodgson GM (1998) The approach of institutional economics. J Econ Lit 36(1):166–192

Hodgson GM (2004) Institutional economics: from Menger and Veblen to Coase and North. In: Davis JB, Marciano A, Runde J. The Elgar companion to economics and philosophy, pp 84–101

Kasper W, Streit ME (1998) Institutional economics: social order and public policy. Edward Elgar, Cheltenham

Keefer P, Knack S (1997) Why don't poor countries catch up? A cross-national test of an institutional explanation. Econ Inq 35(3):590–602

Medema SG, Mercuro N, Samuels WJ (1999) Institutional law and economics. In: Boukaert B, De Geest G (eds) Encyclopaedia of law and economics. Edward Elgar/The University of Ghent, Cheltenham

Nelson RR, Sampat BN (2001) Making sense of institutions as a factor shaping economic performance. Revista Econ Inst 3(5):17–51

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

North DC (1991) Institutions. J Econ Perspect 5(1):97–112

Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. University Press, Cambridge

Persson T, Tabellini G (2003) Economic effects of constitutions. MIT Press, Cambridge

Polanyi K (1944) The great transformation. Rinehart, New York

Rodrik D (2000) Institutions for high-quality growth: what they are and how to acquire them. Stud Comp Int Dev 35(3):3–31

Rutherford M (1994) Institutions in economics: the old and new institutionalism. Cambridge University Press, Cambridge

Sened I (1997) The political institution of private property. Cambridge University Press, Cambridge

Veblen TB (1898) Why is economics not an evolutionary science? Q J Econ 12:373–397

Wade R (1990) Governing the market: economic theory and the role of government in East Asian industrialization. Princeton University Press, Princeton

Weber M ([1904–1911] 1968) Economy and society. Bedminster, New York

Williamson OE (2000) The new institutional economics: taking stock, looking ahead. J Econ Lit 38:595–613

Institutional Review Board

Roberto Ippoliti

Department of Management, University of Turin, Turin, Italy

Scientific Promotion, Hospital of Alessandria – "SS. Antonio e Biagio e Cesare Arrigo", Alessandria, Italy

Abstract

Institutional Review Board (IRB) are independent institutions aimed at approving, monitoring and reviewing human experimentations in order to protect research subjects' rights from the necessity to increase the current medical knowledge.

After a brief historical introduction of humans experimentation, this section presents American and European regulation of this specific institution, as well the main related issues in Law and Economics.

Synonyms

Ethical review board; Independent ethics committee

Definition

Institutional Review Board (IRB) is a committee designated to approve, monitor, and review human experimentations in order to protect research subjects' rights.

IRB and its Evolution

Institutional Review Board (IRB), also known as independent ethics committee or ethical review board, is a committee designated to approve, monitor, and review human experimentations in order to protect research subjects' rights.

The matter of research subjects' rights was first addressed after the end of World War II, when Nazi experiments on prisoners and Jews were discovered. Those researches were motivated by racial and political conflict and the idea of rights for those kept in Nazi concentration camps held no meaning. The subjects were coerced into participating and there was no informed consent about potential related expected and unexpected adverse events (e.g., their death). The Nazi experiments were aimed at supporting the Nazi racial ideology, as well as at improving the survival and rescue of German troops, by testing medical procedures and pharmaceuticals.

After that experience, society felt that it was its duty to prevent research involving people who were used as test subjects not of their own free will and to check the scientific aims of research studies. Thus, The Nuremberg Code of Ethical Human Subjects Research Conduct (1947) was drafted. This was the first international document on human experimentation and research subjects which highlighted the main right of patients: voluntary consent of human subjects involved in clinical trials. Obviously, this information concerns risk-benefit outcomes of experiments, i.e., both expected effectiveness and potential related expected/unexpected adverse events. However, other examples of inhuman clinical research were necessary to develop a systematic review of these studies by independent institutions, for example, the *Tuskegee syphilis experiment*, which was a clinical study conducted by the US Public Health Service to study the natural progression of untreated syphilis, or the *Vipeholm experiments*, which was sponsored by the sugar industry and dentist community, in order to determine whether carbohydrates affected the formation of cavities. In the former study, research subjects were rural American men who thought they were receiving free health care from the US government, even if they were never told they had syphilis nor were they ever treated for it (see Thomas and Quinn (1991)). In the latter case, patients of a Swedish mental hospital were fed large amounts of sweets to provoke dental caries, violating the basics of medical ethics (see Gustafsson et al. 1954).

Over the years, that Code has been followed by other international documents, among which one of the most important was the Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects (1964). This international document was developed by the World Medical Association in 1964 as a means of governing international clinical research. It is mainly a set of guidelines for medical doctors conducting biomedical research involving human subjects, and it includes the principle that research protocols should be reviewed ex ante by an independent committee. According to its recommendations, a technical board must review each clinical trial before enrollment can start. In other words, a third party has the duty to guarantee the main principles established by the aforementioned international code: patient information and scientific validity of protocols. Another international guideline is proposed by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), which is a project that brings together the regulatory authorities of Europe, Japan, and the USA, as well as experts from the pharmaceutical industry, to discuss scientific and technical aspects of pharmaceutical product registration. ICH provides good clinical practice (GCP), i.e., an international quality standard that governments can transpose

into regulations for clinical trials involving human subjects (ICH Topic E 6 (2002)). According to ICH-GCP, an IRB should safeguard the rights, safety, and well-being of research subjects, with a particular attention to trials that may include vulnerable subjects (e.g., pregnant women, children, prisoners). A European standard for clinical trials and patient protection is the ISO 14155, which is valid in the European Union as a harmonized standard good clinical practice (i.e., ISO-GCP).

According to international documents and guidelines, both in Europe and in the USA, a protection system has been developed through the years, based on this Institutional Review Board (IRB). In the United States (USA), Title 45 of the Code of Federal Regulations, part 46 (revised 2009) – which is the reference regulation for US IRB – identifies this board as an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects, with the authority to approve or disapprove research. The purpose of IRB review is to assure that appropriate steps are taken to protect the rights and welfare of humans involved as subjects in the clinical trials, reviewing the research protocol and related materials (e.g., informed consent documents and investigator brochures). According to Directive 2001/20/EC (2001), the European Union (EU) recognizes this board as an independent body in a member state, consisting of health-care professionals and nonmedical members, whose responsibility is to protect the rights, safety, and well-being of human subjects involved in a trial and to provide public assurance of that protection, by, among other things, expressing an opinion on the trial protocol, on the suitability of the investigators and the adequacy of facilities, and on the methods and documents to be used to inform trial subjects and obtain their informed consent.

Considering their composition, FDA's requirements is set in Title 21 of the Code of Federal Regulations, part 56 (revised 2013) – which is an additional regulation for US IRB that oversee clinical trials involved in new drug applications – suggesting that an IRB must have at least five members with enough experience, expertise, and diversity to make an informed decision on whether the research is ethical, informed consent is sufficient, and appropriate safeguards have been put in place. Moreover, the regulation states that if a study that includes vulnerable populations is under investigation, the IRB should have members who are familiar with these groups (e.g., an IRB has to include an advocate for prisoners when considering research that involves them). The European Directive 2001/20/EC (2001) does not specify the composition of each IRB, which is regulated at the national level by each member state.

Several law and economics issues are related to IRBs and their activity. Just to cite a few, Calabresi (1969) focused on the generational conflict, suggesting that IRB should be an expression of the value of research that involves human subjects and how it is necessary to achieve an adequate balancing of present against future lives. Other researchers focus on the effectiveness of the IRB decision-making process, highlighting how these boards have exercised primary oversight responsibility for human research subject, continuing to be often incapable of reviewing complex research protocols effectively (Hoffman 2001; Coleman 2004). Finally, proposing the ideal "Market of Human Experimentation," Ippoliti (2013a, b) analyzes the relation between IRB activity and the transaction costs to achieve an exchange of information for innovation, suggesting the main impact on countries' competitiveness.

Another interesting topic related to the IRB members, which should be deeply analyzed from a law and economic prospective, concerns the undue influence of pharmaceutical companies and the related economic interests involved in the ethical decisions. Just to give an idea of the issue, Campbell et al. (2006) analyze the relation between IRB members and industry, suggesting that 36 % of these members had had at least one relationship with for-profit institutions in the past year. Moreover, authors denote that of the respondents, 85.5 % said they never thought that the relationships that another IRB member had with the industry affected his or her

IRB-related decisions in an inappropriate way. The economic undue influence is even more important considering the US market, where some IRB reviews are conducted by *for-profit* organizations – i.e., commercial IRBs (Emanuel et al. 2006).

Cross-References

► Human Experimentation

References

American regulation: title 45 of the Code of Federal Regulations – part 46 (2009) Title 21 of the Code of Federal Regulations– part 56 (revised 2013)

Calabresi G (1969) Reflection on medical experimentation in human. Daedalus 98(2):387–405

Campbell EG, Weissman JS, Vogeli C, Clarridge BR, Abraham M, Marder JE, Koski G (2006) Financial relationships between institutional review board members and industry. N Engl J Med 355:2321–2329. https://doi.org/10.1056/NEJMsa061457

Coleman CH (2004) Rationalizing risk assessment in human subject research. Arizona Law Rev 46(1):1–51

Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects (1964)

European Union regulation: Directive 2001/20/EC (2001)

Emanuel EJ, Lemmens T, Elliot C (2006) Should society allow research ethics boards to be run as for-profit enterprises? PLoS Med 3(7):e309. https://doi.org/10.1371/journal.pmed.0030309

Gustafsson BE, Quensel CE, Lanke LS, Lundqvist C, Grahnen H, Bonow BE, Krasse B (1954) The Vipeholm dental caries study; the effect of different levels of carbohydrate intake on caries activity in 436 individuals observed for five years. Acta Odontol Scand 11(3–4):232–264

Hoffman S (2001) Continued concern: human subject protection, the institutional review board, and continuing review. Tenn Law Rev. 2001 Summer 68(4):725–70

ICH Topic E 6 (R1) Guideline for Good Clinical Practice (2002)

Ippoliti R (2013a) Economic efficiency of countries' clinical review processes and competitiveness on the market of human experimentation. Value Health 16:148–154

Ippoliti R (2013b) The market of human experimentation. Eur J Law Econ 35(1):61–85

The Nuremberg Code of Ethical Human Subjects Research Conduct (1947)

Thomas SB, Quinn SC (1991) The Tuskegee syphilis study, 1932–1972: implications for HIV education and AIDS risk programs in the black community. Am J Public Health 81:1503

Insurance Market Failures

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

Abstract

Insurance market is characterized by failures that impose particular negative consequences; given the failures, different remedies may improve the market outcome. On one hand, the insurance market is characterized by asymmetric information, i.e. moral hazard and adverse selection, and to correct the consequent severe market failures, monitoring and risk classification can be implemented. On the other hand, the insolvency issue: given the enormous amounts of funds in the hands of insurance companies, their default would have an extreme impact, and regulation is necessary to guarantee the payback for policyholders and beneficiaries.

Definition

Insurance is an instrument to give protection against the risk resulting from various perils or hazards, such as health risks, invalidity or death, accidents, unemployment, theft, fire, and many more. Contracts are offered by insurance companies providing risk-sharing mechanisms that allow their customers to replace these risks. Insurance market is characterized by failures that impose particular negative consequences on one or both market sides: given the failures, different remedies may improve the market outcome.

Introduction

Insurance plays three economic functions: (i) the transfer of risk from a risk-averse individual to the risk-neutral insurer, (ii) the pooling of risk so that the "uncertainty" of each insured becomes the

"certainty" of the insurance companies that this risk will occur to a percentage of their customers, and (iii) the allocation of risk for which each insured pays a price that should reflect the risk he contributes (Abraham 1995).

For these three reasons, insurance contracts increase social welfare while at the same time induce people to have a preventive behavior and contribute to internalize damage. At a macroeconomic level, decreasing the economic effects of risks, insurance encourages companies to operate in risky sector and to make investments that they would not make otherwise. Meanwhile, life insurance plays a role as a long-term investment and savings instrument (Shavell 2000).

Asymmetric Information

The insurance market is characterized by the fundamental problem of bilateral asymmetric information. On one hand, individuals do not have complete information in understanding complicated insurance contracts and lack the ability to assess the adequacy of premium to risk. On the other hand, insurers suffer from lack of information regarding the risk characteristics of individuals. This second asymmetry generates the two phenomena of moral hazard and adverse selection.

Moral hazard (hidden action) depends on the insurers' impossibility to perfectly know the extent their customers' behavior may affect the occurrence and/or the dimension of the loss.

To be precise, the term moral hazard refers to at least two different situations in which the insured's behavior can affect the probability of the various outcomes: (i) situations when insurance may induce greater use of a service by an insured individual or cause the insured to exercise less care and (ii) situations when an insured individual purposely causes harm or otherwise falsifies loss in order to collect insurance benefits or to inflate the loss.

In the case of moral hazard, the insured's behavior changes, and the insurer is unable to either predict this change in advance or to prevent it by exempting such behavior from the insurance contract coverage (Shavell 1979).

In fact, after signing an insurance contract, the insured may have less incentive to act carefully or take preventive measures, influencing both the damage probability and/or the loss dimension.

Moral hazard, even if severe, does not cause a complete breakdown of markets, but it raises the cost of insurance and, consequently, reduces the degree of insurance coverage negatively affecting market outcomes (Tennyson and Warfel 2009).

In this case, a remedy is monitoring the insured's behavior: ex ante the loss occurrence, to monitor the level of care in preventing the loss, insofar as the insured's behavior has any influence over the risk; ex post, to monitor the amount of claim when loss occurs, beyond the services the claimant would purchase if not insured and assuming the insured individual can influence the magnitude of the claim. Also incentive schemes linking the price of insurance to observed past behavior (e.g., bonus/malus systems) can be used as devices to contain this problem (Derrig 2002).

Adverse selection (hidden information) refers to the inability of insurers to observe risk characteristics of their customers, leading to offer a contract based on the average risk of the entire group of customers. In this case, more high-risk individuals purchase insurance; higher payouts by insurance companies force them to raise rates which, in turn, makes the insurance less attractive to low-risk individuals.

As a consequence, this may reduce the stability of the market equilibrium, and the market may completely break down, such as the famous "market for lemons" (Akerlof 1970).

In the case of adverse selection, a remedy would be to use statistical data to separate different categories of risky individuals by classification instrument.

Theoretically in determining the premium to be charged, insurers should estimate the expected losses for each individual being insured. But given the informational asymmetry, the insurance companies apply risk classification trying to group the individuals in such a way that those with a similar loss probability are charged the

same rate. The risk classification systems are clearly supported by statistical data showing differences in the event rate in different groups (Porrini 2015).

In practice, insurers have to identify risks that are independent, uncorrelated, and equally valued and to aggregate them in order to reduce the total risk of the set. An efficient risk classification reduces adverse selection, because it makes insurance more attractive to the low-risk individuals.

Classifying insurance risks increases the efficiency of contracting in terms of asymmetric information. However, the benefits are conditional on general principles, such as the nondiscrimination, and generally to consumer protection issue.

Insolvency

Generally, the default of a company generates economic damages, first to shareholders, but in many cases also to the customers. Particularly, in insurance market, the insured individuals may lose future benefits and insurance coverage with possibly precarious economic situation in many cases and imposing to rely on a public coverage of these losses.

Moreover, this is reinforced by the consequences of the so-called inversion of the production cycle that comes from the fact that insurance services are only delivered after they are purchased and in many cases years later. This creates the necessity to monitor the financial condition and solvency of insurance companies over an extended period of time.

This feature leads potentially to insufficient capitalization and suboptimal solvency levels, by giving insurers scope for hiding poor underwriting and under-reserving, and these are motivations for government interventions aimed at monitoring to improve management discipline.

Regulation for solvency dates to the nineteenth century, when insurance insolvencies in the USA and Europe led to the establishment of state regulatory authorities. Given the social role and the involvement of insurance in the systemic risk issue (Faure and Hartlief 2003), solvency becomes the primary focus of insurance regulation worldwide. Regulatory tools include risk-based capital requirements, electronic auditing of accounts, and a wide variety of limits on the ways that companies can invest the funds held in reserve to pay claims.

Moreover, regulation can be used to steer capital into preferred fields, given that insurance is an institution for storing and accumulating capital, competing with banking and securities firms. Although banking, insurance, and securities have traditionally been subject to different regulatory regimes, there is recently a "convergence" in the financial services marketplace that places great strain on the existing regulatory institutions (i.e., the diffusion of the business model of bank insurance).

The most common instrument of regulation for solvency consists of technical provisions that correspond to the amount required by the insurer to fulfil its insurance obligations and settle all commitments to policyholders arising over the lifetime of the portfolio. Technical provisions can be divided into those that cover claims from insurance events which have already taken place at the date of reporting and those that should cover losses from insurance events which will take place in the future.

Most countries supplement the above requirements by regulating the portfolio choices of insurance firms with the aim to ensure that insurers invest and hold adequate and appropriate assets to cover capital requirements and technical provisions.

The main focus of numerous national regulation is to ensure that insurance companies are able to honor their payment obligations in a continuous way, the most important instrument being that of a generalized capital and reserve rules, possibly supplemented by additional supervisory rules, such as investment restrictions and provisions of regular inspection by supervisory authorities.

Conclusion

Because the insurance business has become highly important to society's development, it is

relevant to find remedies to the failures that can impede a correct functioning of the market.

On one hand, the insurance market is characterized by fundamental problems of asymmetric information. Moral hazard and adverse selection play a central role in the insurance market, and to correct the consequent severe market failures, monitoring and risk classification can be implemented.

On the other hand, the insolvency issue is justified by the enormous amounts of funds and investments in the hands of insurance companies; as such, their default would have an extreme impact, and regulation is necessary to guarantee the payback for policyholders and beneficiaries.

Reference

Abraham K (1995) Efficiency and fairness in insurance risk classification. Virginia Law Rev 71:403–451

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

Derrig RA (2002) Insurance Fraud. J Risk Insu 69:271–287

Faure M, Hartlief T (2003) Insurance and expanding systemic risks. OECD, Paris

Porrini D (2015) Risk classification efficiency and the insurance market regulation. Risks 4:445–454

Shavell S (1979) On moral hazard and insurance. Q J Econ 93:541–562

Shavell S (2000) On the social function and regulation of liability insurance. Geneva Pap Risk Insur 25:166–179

Tennyson S, Warfel WJ (2009) Law and economics of first party insurance bad faith liability. Conn Insur Law J 16(1):203–242

Intellectual Property: Economic Justification

Dennis W. K. Khong
Centre for Law and Technology, Faculty of Law, Multimedia University, Melaka, Malaysia

Definition

Economic justification for intellectual property means the economic reason for establishing and supporting a system of intellectual property laws. It begins with the characteristics of information as public goods (Arrow 1962): non-rivalry in consumption and non-excludability. Intellectual property rights confer upon the right holders an exclusive right to legally compel users to buy a genuine product from the right holders or to obtain a license from them. Intellectual property rights are to prevent a market failure due to free-riding activities of non-payers. In general, intellectual property rights can be grouped into two families according to the function of the information therein: information as a good and information as a signal. Examples of intellectual property rights in the form of information as a good are copyright, patents, industrial designs, layout-designs of integrated circuits, and confidential information, while examples of intellectual proprietary rights in the form of information as a signal are trademarks, the tort of passing off and unfair competition, and geographical indications.

Introduction

This essay examines the economic justification for intellectual property. The economic justification begins with the twin characteristics of information as public goods: non-rivalry in consumption and non-excludability. Market failure of free riding results from the non-excludability of public goods. Granting intellectual property rights is a solution to solving this public goods problem by conferring upon the creators of new information an exclusive right. In general, intellectual property rights can be grouped into two families according to the function of the information therein: information as a good and information as a signal. Macroeconomic justifications are also discussed. Exceptions and limitations to intellectual property rights are seen as solutions to curb the ill effect of a monopoly right granted by intellectual property. Finally we examine the use of the economic justification as a yardstick to assess proposals for new forms of intellectual property rights.

Intellectual property rights are intangible rights in information. The World Trade Organization's

Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) requires member states to have laws protecting the seven most common types of intellectual property rights, namely, copyright, trademarks, geographical indications, industrial designs, patents, layout-designs (topologies) of integrated circuits, and undisclosed (or confidential) information. In some countries such as in the European Union, database rights are also available.

Since intellectual property rights essentially create a limited form of monopoly, it is necessary to examine the economic justification for them.

Information as a Public Good

The economic justification for intellectual property rights begins with the characteristics of information. Information is said to possess the twin characteristics of public goods (Arrow 1962): non-rivalry in consumption and non-excludability (Sidgwick 1887; Pigou 1923; Samuelson 1954).

Non-rivalry in consumption refers to the characteristic of goods being not exhaustible by consumption. In other words, a public good does not decrease in both quantity and quality as more users consume the good or when the amount of consumption of that good increases. Typically, the cost structure of a public good is said to have a fixed cost, of which is usually considered high, and zero marginal cost, that is, the cost of additional usage is theoretically zero. This quality of non-rivalry in consumption is a good characteristic because it makes the good inexhaustible for consumption and is thus an infinite good.

The second characteristic of public goods is non-excludability. Non-excludability means that once the public goods have been released or made available to the public, it is extremely difficult, if not impossible, to exclude non-payers from consuming or enjoying the benefits, or to compel users to voluntarily pay for the public goods, because it can be enjoyed without payment. It is this characteristic which is the root cause of the public goods problem. Pigou (1923) explains non-excludability as "instances in which marginal private net product [falling] short of marginal social net product, because incidental services are performed to third parties from whom it is technically difficult to exact payment." Non-excludability is a bad characteristic because it often leads to the problem of free riding, i.e., consumers enjoying its benefits without paying for its costs.

Using the above two characteristics as the defining criteria for public goods, it can be shown that information can be classified as public goods. By information, we do not confine it to bits of information but also include knowledge and electronic data. At the same time, we exclude from the definition of information its medium of carrier or the physical manifestation of information. For example, a novel consists of the physical medium, namely, paper and ink, and the information. Paper and ink have a cost component and by definition are rivalrous in consumption. On the other hand, the informational content of a novel, the storyline, and the plot and character development are intangible in nature and are non-rivalrous in consumption. One person's knowledge and appreciation of the novel's content do not hinder another person's knowledge and appreciation of the same.

Likewise a motor vehicle such as a car contains both the physical embodiment of information and its informational content. The metal, rubber, and other materials used in manufacturing a car are merely the physical embodiment of the informational content of the design and technology behind the car. Although in this case not many people would be able to fully comprehend the technical details in the design of a car, the same analysis holds in that this information does not decrease in both quantity and quality in the hands of a person who could fully comprehend it. Moreover it is non-excludable because short of never selling to the public the said car, once the car is available in the market, a technically competent competitor might be able to reverse-engineer the design of the car without paying the original inventor, unless the force of law compels him to do so.

In the case of digital music, the music files have no physical embodiment short of some electromagnetic configurations. Digital files can be

easily copied at negligible or no reproduction cost. And once the music is made available on a digital network such as the Internet, it is extremely difficult or impossible to limit access to it to payers only. Thus digital music also exhibits the characteristics of public goods.

Market Failures

Information as public goods suffers from market failure due to free riding. Free riding or freeloading is defined as enjoying a benefit without paying for its cost. Two types of market failures may occur as a result. First is sub-optimal level of investment in the creation of information because of not taking into account the social benefits of information. The second is no or little investment in the creation of new information due to the threat of free riding.

To illustrate the first, we can use the example of computer software. A lone programmer who has no means of preventing free riding if he releases a computer program he has written will typically not be able to create a software package as comprehensive and well-designed as a piece of commercial software done by a software firm, because to develop commercial software would require a more extensive manpower and financial investment than could be shouldered by a typical lone programmer. So in the absence of the ability to obtain property rights in his software to the exclusion of non-paying customers, the lone programmer would probably develop his software in a sufficiently simple way to satisfy his own personal requirement. In other words, investment of effort by the programmer would be to the extent of marginal cost equals to marginal private benefit. This can be considered as a sub-optimal level of investment as optimality would require the programmer to take into account the social benefit of his creation and thus expand the programming effort to the extent of marginal cost equals to marginal social benefit.

The second form of market failure exists in the form of sub-optimal level of new information being created. This market failure occurs when potential copiers and users wait out for someone else to create the new information. Since everyone is waiting for someone else to take the first step in creating new information, no one or few will end up doing the work of creation. This is akin to the case of prisoner's dilemma (Gordon 1992) where a cooperative strategy would be the optimal outcome, but since everyone's dominant strategy is to wait to free ride, the resulting Nash equilibrium is that all players will choose the free-riding strategy and no new information is created.

Solutions to Public Goods Problem

The traditional solution to the public goods problem is state provision from taxation as a means of funding. Both state provision and taxation make sense in relation to public goods, the benefits of which are enjoyed by the population at large, e.g., national defense, public health, and machineries of the government. However, it is not given that public goods must be provided by the state alone nor to be funded solely through a system of taxation. Non-state entities can be paid to provide public goods, or funding can be made through some form of nonvoluntary payment imposed on users. For example, Coase (1974) reminded us that in the 19th century, some lighthouses in England were operated by private parties. Passing ships were required to pay compulsory light dues.

In the same way, instead of relying on the state to create or fund the creation of new information, intellectual property rights as a form of property rights which allow creators to collect a user license fee become the preferred way to solve the problem of information as public goods. Indeed, Pigou (1923) in his *magnum opus*, *The Economics of Welfare*, discussed the public goods problem of scientific knowledge and the role of patent law to mitigate the problem of public goods.

There are several advantages to using intellectual property rights as a solution. First is that there will be less cross-subsiding by nonusers for the benefits of users. Using a system of intellectual property rights would mean that only users of those intellectual properties will have to pay license fees or purchase original copies.

Another advantage is that different creators will find their own profit-maximizing strategy based on what they perceive as the market demand, their own abilities, and preferences. So there is no difficulty of requiring central planning and the need for the state to decide on what information to create for its people. It also allows minority creators outside the mainstream society to cater to the demands of the members of the society at the fringe. In other words, intellectual property rights, to a certain extent, promote liberty and freedom of speech and expression.

On the other hand, intellectual property rights do not preclude state funding in the creation of new information. Indeed it is essential for the state to fund research both for its own use, such as for education and policy, and for public consumption, such as in the case of scientific research.

Intellectual property rights as a form of intangible property rights have another advantage in that it is transferable. This means that the potential value in an intellectual property can be realized by the owner selling or exclusively licensing it to another. Creators may choose this route in order to focus on the act of creation and leave the business of commodification to others.

Two Families of Intellectual Property Rights

Traditional legal treatment of intellectual property law tends to lump all different types of intellectual property rights into a single conceptual basket called "intellectual property." Rather, all these different types of intellectual property rights are treated as discrete forms with no apparent similarities among them. Although this approach has its advantage for its simplicity, it nevertheless obscures the true nature and functions of different types of intellectual property rights. In general, intellectual property rights can be grouped into two families according to the function of the information therein. The first is information as a good, and the second is information as a signal.

Examples of intellectual property rights in the form of information as a good are copyright, patents, industrial designs, layout-designs of integrated circuits, and confidential information. The commonality among these types of intellectual property rights is that the information so protected is an object that can be consumed to increase one's utility, such as a story, a picture, a piece of music or art, an invention, a design, or some proprietary knowledge.

Information as a signal on the other hand covers intellectual property rights such as trademarks, passing off, and geographical indications. The information therein is not an object of consumption per se and cannot be enjoyed directly. Instead it helps us to use the market better by providing indications as to the sources and quality of goods and services associated with those intellectual property rights. Typically, information as a signal comes in the form of a brand, a logo, or an image of a personality, which may convey the origin or endorsement of a product.

These two families of information suffer from different forms of free riding and thus merit separate investigation.

Free Riding in Information as a Good

Free riding in information as a good happens when a free-rider, without the right holder's consent, uses or reproduces an intellectual property. Thus, intellectual property rights are used to correct this market failure by granting to its creators and owners an exclusive or monopoly right to exploit the information so created (Plant 1934). Typically this occurs in two ways.

The first way of exploitation is in the form of products embodying the intellectual property, such as books, digital content, and technological inventions. As intellectual property rights give an exclusive right to the owner, the owner can then sell his products at a profit-maximizing monopoly price which is higher than their marginal costs of production. It is this higher than market competitive price which hopefully will bring excess profit sufficient to cover the high cost of the creation of the proprietary information.

The second way of exploitation is by way of licensing so that profit is made by allowing others to exploit the protected intellectual property

instead of selling products by the intellectual property owners themselves.

Free Riding in Information as a Signal

Intellectual property rights are also used to correct the market failure from free riding of information as a signal. Examples of this type of intellectual proprietary rights are trademarks, the tort of passing off in common law jurisdictions (or unfair competition in others), and geographical indications (Ritzert 2009).

Traditionally trademarks and the tort of passing off are said to protect the trade origin of or goodwill associated with the products. What this means is that the protected trademark serves as a link between the product and its originator, i.e., the trademark owner.

Businesses apply trademarks to products and services in order to distinguish the products and services originating from them from those of competitors. In a competitive market, or even in a duopoly, businesses increase profit by selling more. One of the ways of doing so is to attract and retain customers by offering a superior product compare to the competitors'. Even for market segment with low-quality products, businesses retain customers by offering products of a consistent quality. Alternatively, businesses attract customers by way of advertising in order to get potential customers familiar with their products through an advertising campaign.

Both strategies of superior or consistent product quality and advertising are not costless. A free-riding competitor may steal a trademark owner's market share by selling counterfeit products bearing the right holder's trademark without incurring the above costs. In fact, counterfeiters can do better in the short run by selling cheap, low-quality counterfeit products if quality is costly.

Trademark rights therefore grant the trademark owners an exclusive right to use his trademark for the classes of products or services it is registered to. With this exclusive right, the trademark owner can legally prevent counterfeits from using his trademark by way of criminal and civil enforcements.

Another economic justification for trademark protection is to prevent consumer confusion. Supposing that a consumer values a genuine product highly and pays a high price for this product, but the product sold is a low-quality counterfeit, it would mean that the consumer has suffered a disutility. Trademark protection thus can be justified to prevent this form of economic transactions.

Therefore the economic functions of trademarks as intellectual property rights are to encourage optimal investment in advertising, optimal investment in product quality on the part of producers, and optimal consumption on the part of consumers (Landes and Posner 1987; Economides 1988). Without trademark rights, counterfeiters will reduce a trademark owner's incentive to promote his products bearing his trademark because part of the potential sale will be diverted to the counterfeiters as consumers are misled into buying counterfeit goods.

In addition, low-quality counterfeit goods in the market will jeopardize the trademark owner's effort in ensuring high and consistent quality of products bearing his trademark. Like in the market for lemons (Akerlof 1970), low-quality counterfeit goods will crowd out high-quality goods by the trademark owner as consumers could not differentiate the quality of counterfeit goods from genuine goods by way of observation. Under the condition of uncertainty in the quality, consumers will treat the market of the trademark goods as having an average quality and will have a lower willingness to pay, leading to a low equilibrium price. In the end, without trademark protection, low-quality counterfeit goods will crowd out high-quality genuine goods, making high-quality goods unavailable in the market.

Macroeconomic Justifications

Intellectual property rights are also justified from a macroeconomic perspective, especially in relation to intellectual property rights being incorporated into international trade treaties.

One argument, in the context of technology transfer, is that foreign businesses are reluctant to introduce or bring in new technology or

know-how to a country which does not have sufficiently broad and strong intellectual property rights protection. Intellectual property protection is seen as a pull factor for foreign investment, be it technological transfer or sending the design of a product to be manufactured in the receiving country (Mansfield 1995; Fink and Maskus 2005). Hence, intellectual property is argued as an exogenous growth factor.

Another macroeconomic justification builds on the relationship between intellectual property rights and economic growth (Gould and Gruben 1996). For example, businesses may earn higher profits with trademark protection. In addition, businesses are seen as having strong incentive to develop new knowledge and technology with copyright and patent protections (Sweet and Maggio 2015). Nevertheless, evidence of positive correlation between intellectual property rights and economic growth is ambiguous in many sectors (Aziz 2003; Yueh 2007).

Economic Justifications Versus Motivation

Economic justifications for intellectual property right are theoretical reasons or, at most, plausible empirical support for a system of intellectual property law. Empirical studies have uncovered other reasons and personal motivations for creators to create new works and inventions. For example, it was found that computer programmers voluntarily contribute to open-source projects because of identification with a group, pragmatic motivation to improve one's software or career, social and political motivation to support a movement, and enjoyment derived from the process itself (Hertel et al. 2003).

Exceptions and Limitations to Intellectual Property Rights

Although propertization through intellectual property rights is the preferred method of solving the market failure in information problem, it on the other hand creates a different form of market failure, i.e., monopoly from exclusive right. In order to contain and limit the ill effects of monopoly power from intellectual property rights, various legal strategies are used, some of which are discussed below.

The first strategy is to control the breath of the intellectual property rights, i.e., the extent it could prevent other similar creations from being lawfully used. Limiting the breath of protection will facilitate competitors' ability to produce close but differentiated substitutes in the marketplace and thereby foster a market characterized by monopolistic competition (Chamberlin 1933) or imperfect competition (Robinson 1933). For example, copyright law protects expressions and not ideas. So general ideas of fictional stories are not protected, and authors may write fictional stories of the same genres as long as the detailed plots and character names are not colorably similar. Similarly, different publishers may produce their own language dictionaries containing substantially similar word list but with their own definitions.

The second strategy is to limit the term or duration of protection. Almost all forms of intellectual property rights, except for trademarks and confidential information, have limited term of protection. Limiting the term of protection has the beneficial effect of curbing the static inefficiency from monopoly power and allows users and other producers to have unlimited use of the intellectual property when it lapses into the public domain upon the expiry of its term (Walterscheid 2000). Although trademarks may be renewed indefinitely, registration may be revoked for nonuse, so as to prevent useful trademarks from being hoarded by first-moving trademark owners. Undisclosed confidential information gets protection so long as the information is kept secret and not disclosed to the public. It loses protection immediately when an independent party rediscovers the confidential information through independent effort or through reverse-engineering of a publicly available product.

The third strategy is to create situational exceptions to protection. An economic justification for such an exception is that the transaction cost of licensing for the use of the intellectual property would outweigh the benefit of using the

intellectual property. Potential users will be deterred from using an intellectual property if the only legal route is through licensing. Example of such an exception can be found in copyright's fair use doctrine (Gordon 1982; Gordon 2002). In jurisdictions with statutory exceptions in copyright law, situational exceptions for disadvantaged groups such as the accessibility exception for the disabled can be justified on the ground that not all publishers are willing or interested in selling products which caters to a small market, and thus allowing the disabled to use technological solutions, such as text-to-speech software, to solve their accessibility problem is socially desirable.

The fourth strategy is to allow compulsory licensing in exceptional circumstances. Compulsory licensing can be seen as a form of liability rule protection of intellectual property rights (Reichman 2009). When compulsory licensing is evoked, the intellectual property owner has no power to set the price but has to instead accept the price set by an authoritative body, which in some circumstances is below a monopoly price. It also dispenses with the potentially protracted negotiation process in a national emergency. Compulsory licensing provision can be found in patent law which allows the state to manufacture essential medicines when patent holders refuse to grant a license at a reasonable price. Despite its unpopularity, compulsory licensing is rarely used and most often acts as a threat to force patent holders to lower their price of medicine.

Economic Justifications as a Yardstick

Assuming that the proper justification for intellectual property rights is economic in nature, i.e., to solve a market failure problem, then this economic justification can also be used as a yardstick to assess whether new forms of information should be conferred an intellectual property rights protection.

Certain countries and communities are pressing the World Intellectual Property Organization (WIPO) to recognize and protect traditional knowledge. Unlike patent which protects only new inventions, traditional knowledge protection, as currently defined by WIPO, protects "knowledge, know-how, skills and practices that are developed, sustained and passed on from generation to generation within a community" (Matsui 2015). It appears that since traditional knowledge has been around the relevant communities for a long time, it would not be capable of patent protection because it would not satisfy the novelty requirement of patent law. The fact that traditional knowledge already exists means that there is no lack of an incentive to create problem. Instead the usual justification for protecting traditional knowledge is to reward or compensate the community for developing and preserving the traditional knowledge. However, this justification is actually of a distributional concern and not of an efficiency concern. In other words, traditional knowledge protection is merely a mechanism to transfer some wealth from an often wealthy first-world exploiter to a usually, but not necessarily, poorer and less-developed community.

Another application of the economic justification can be seen in assessing the arguments for retroactive extension of intellectual property term. Opponents of retroactive term extension argue that since retroactive term extension does not incentivize the creation of new intellectual property, it is not a good legal policy as it contradicts the basis for intellectual property rights which is to solve the market failure problem of information (Akerlof et al. 2002).

Cross-References

- ▶ Copyright
- ▶ Creative Commons and Culture
- ▶ Digital Piracy
- ▶ European Patent System
- ▶ Geographical Indications
- ▶ Lighthouses
- ▶ Patent Litigation
- ▶ Patent Opposition
- ▶ Public Goods
- ▶ Trade Secrets Law
- ▶ Trademark Dilution

- ▶ Trademarks and the Economic Dimensions of Trademark Law in Europe and Beyond
- ▶ TRIPS Agreement

References

Akerlof GA (1970) The market for "lemons": quality uncertainty and the market mechanism. Quarterly J Econ 84:488–500

Akerlof GA, Arrow KJ, Bresnahan TF, Buchanan JM, Coase RH, Cohen LR, Friedman M, Green JR, Hahn RW, Hazlett TW, Hemphill CS, Litan RE, Noll RG, Schmalensee R, Shavell S, Varian HR, Zechkauser RJ (2002) Brief as amici curiae in support of petitioners in Eldred v Ashcroft 537 US 186 (2003) (No 01–618)

Arrow K (1962) Economic welfare and allocation of resources for inventions. In: Universities-National Bureau Committee for Economic Research C on EG of the SSRC (ed) The rate and direction of inventive activity: Economic and social factors. Princeton, New Jersey University Press, pp 609–626

Aziz S (2003) Linking intellectual property rights in developing countries with research and development, technology transfer, and foreign direct investment policy: a case study of Egypt's pharmaceutical industry. ILSA J Int Comp Law 10:1–34

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

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

Economides NS (1988) The economics of trademarks. Trademark Report 78:523–539

Fink C, Maskus KE (eds) (2005) Intellectual property and development: lessons from recent economic research. World Bank and Oxford University Press, Washington, DC

Gordon WJ (1992) Asymmetric market failure and prisoner's dilemma in intellectual property. Univ Dayt Law Rev 17:853–870

Gordon WJ (1982) Fair use as market failure: a structural and economic analysis of the Betamax case and its predecessors. Columbia Law Rev 82:1600–1657.

Gordon WJ (2002) Excuse and justification in the law of fair use: transaction costs have always been part of the story. J Copyr Soc USA 50:149–198

Gould DM, Gruben WC (1996) The role of intellectual property rights in economic growth. J Dev Econ 48:323–350

Hertel G, Niedner S, Herrmann S (2003) Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. Res Policy 32:1159–1177

Landes WM, Posner RA (1987) Trademark law: an economic perspective. J Law Econ 30:265–309

Mansfield E (1995) Intellectual property protection, direct investment, and technology transfer: Germany, Japan, and the United States. Washington, DC

Matsui K (2015) Problems of defining and validating traditional knowledge: a historical approach. Int Indig Policy J 6:art 2

Pigou AC (1923) The economics of welfare, 3rd edn. Macmillan, London

Plant A (1934) The economic aspects of copyright in books. Economica 1:167–195

Reichman JH (2009) Compulsory licensing of patented pharmaceutical inventions: evaluating the options. J Law, Medicine & Ethics 37:247–263

Ritzert M (2009) Champagne is from champagne: an economic justification for extending trademark-level protection to wine-related geographical indicators. AIPLA Q J 37:191–225

Robinson J (1933) The economics of imperfect competition. Macmillan, London

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

Sidgwick H (1887) The principles of political economy. Macmillan, London

Sweet CM, Maggio DSE (2015) Do stronger intellectual property rights increase innovation? World Dev 66:665–677

Walterscheid EC (2000) Defining the patent and copyright term: term limits and the intellectual property clause. J Intellect Prop Law 7:315–394

Yueh LY (2007) Global intellectual property rights and economic growth. Northwest J Technol Intellect Prop 5:436–448

Internal Motivations

- ▶ Intrinsic and Extrinsic Motivation

International Litigation and Arbitration

S. I. Strong
University of Missouri School of Law, Columbia, MO, USA

Abstract

The world of international dispute resolution has changed significantly in the last ten to fifteen years. International actors now must consider a wide array of options regarding

how, when and where their legal disputes will be resolved. This entry discusses the various means of resolving international legal disputes and outlines how law and economics analysis has helped rationalize an increasingly chaotic field of law. In so doing, the discussion considers four core areas of concern: the effectiveness of international dispute resolution; competition between and within different dispute resolution mechanisms; choice of substantive and procedural law and conflict of laws; and third-party funding.

Definition

The field of international dispute resolution is densely populated and includes a variety of types of litigation and arbitration. Not only do these mechanisms differ from each other; they also vary significantly from domestic forms of litigation and arbitration. Scholars must take these distinctions into account if their research is to be credible.

International litigation can refer to the judicial resolution of claims in either national court (sometimes called *transnational litigation*) or international court. An international court may operate regionally (as does the European Court of Human Rights) or globally (as does the International Court of Justice). Disputes heard in international courts are governed by public international law. Although claims arising under public international law may also be heard in national courts, international litigation in national courts typically involves matters of private international law. As a result, international litigation in national courts is often resolved pursuant to domestic legal principles, with the only "international" element arising as a result of the nationalities of the parties.

International commercial arbitration refers to final and binding adjudication of a dispute by a private, neutral decision-maker (i.e., a single arbitrator or a three-person arbitral tribunal). The definition of "international" varies according to national legislation but typically involves parties from different jurisdictions or parties from the same jurisdiction arbitrating in a foreign country. The definition of "commercial" also varies according to national law, although a number of legal systems simply require some sort of economic effect.

International commercial arbitration is characterized as a "private" form of dispute resolution because it only arises upon the consent of the parties. However, international commercial arbitration also includes a number of public features. For example, the international commercial arbitration regime is built upon an intricate web of international treaties and national legislation designed to support the enforcement of both arbitration agreements and arbitration awards across national borders. These legal instruments have been construed by courts from around the world on numerous occasions, and the data has been collected and published by public and private entities for over 50 years (Strong 2009). The United Nations Commission on International Trade Law (UNCITRAL) is one of the more important public providers, having created a freely accessible online source (CLOUT, or Case Law on UNCITRAL Texts) with a wide variety of national court decisions concerning the various model laws and conventions promulgated by UNCITRAL.

As a matter of practice, international commercial arbitration bears relatively little resemblance to domestic forms of arbitration, including consumer, employment, labor, or final offer arbitration. These latter mechanisms are often extremely informal and typically feature a decreased emphasis on legal authorities and argument. International commercial arbitration, on the other hand, is a very sophisticated and highly legalistic procedure that resembles complex commercial litigation much more than it does domestic arbitration. Awards rendered in international commercial arbitration are usually fully reasoned and are often published in denatured (anonymized) form, thus allowing the international legal community to evaluate and predict arbitral behavior despite the confidentiality of the proceedings themselves (Strong 2009).

Investment arbitration (also referred to as *investor-state arbitration* or *treaty-based arbitration*)

refers to arbitral proceedings that arise pursuant to a bilateral investment treaty (BIT), a multilateral investment treaty (MIT), a free trade agreement (FTA), or an investment protection agreement (IPA). Although *interstate arbitration* (meaning arbitration between two nation-states) may also be initiated under many of these instruments, such proceedings seldom occur. Instead, it is more likely that an individual investor will file proceedings seeking redress from injuries allegedly caused by the actions of the host state.

Investment arbitration can be distinguished from *contract-based arbitration* (which includes both international commercial arbitration and domestic forms of arbitration) in several ways. For example, consent to investment arbitration does not rely on contractual privity between the investor and the respondent state but instead arises as a result of a standing offer of arbitration from the respondent state (as reflected in the relevant treaty or other international agreement) to the individual investor. Investment arbitration also differs from international commercial arbitration in that the former primarily addresses questions of public international law. As a result, investment arbitration has been analogized to international courts in some key regards (Born 2012).

The procedural sophistication and formality of investment arbitration is similar to that of international commercial arbitration. In fact, investment arbitration is sometimes governed by procedural rules originally developed for use in international commercial arbitration. However, the quasi-public nature of investment arbitration has permitted or required the introduction of a number of procedural features not seen in other types of arbitration, including the possibility of third-party participation through the filing of *amicus*-type submissions. The need for transparency in investment arbitration has also led to the routine publication of investment awards, often in their original rather than denatured form. Investment awards are fully reasoned and may include dissenting opinions.

Although litigation and arbitration are the best-known methods of international dispute resolution, *international mediation* has received an increasing amount of attention in recent years in both the international commercial and investment contexts. The surge of interest in international mediation is somewhat surprising, given the long-standing availability of *international conciliation*. However, international conciliation has never been as popular as international arbitration as a means of resolving cross-border conflicts.

Although considerable debate exists as to whether conciliation and mediation are identical in all regards, scholars agree that both mechanisms are nonbinding (unless and until the parties execute a binding settlement agreement) and cooperative rather than adjudicative. Like arbitration, mediation and conciliation are private dispute resolution mechanisms that arise through the consent of the parties. Although a full analysis of the law and economics literature concerning international mediation and conciliation is beyond the scope of the current discussion, a number of commentators have suggested that mediation and conciliation provide a more effective and cost-efficient means of resolving international disputes (Spain 2010; Welsh and Schneider 2013). Further research in this field would appear to be beneficial and could possibly build on the large and growing number of empirical and economic analyses concerning the efficacy of mediation and settlement in the domestic context. However, future work would need to specifically address the effectiveness of these mechanisms in the international realm.

Another area of inquiry that is not addressed in this entry involves interstate trade disputes, such as those heard by the World Trade Organization (WTO) or arising under certain FTAs. Although law and economics scholars have considered these types of concerns on occasion (Guzman and Simmons 2002; Symposium 2012), these matters are traditionally not considered under the rubric of international litigation or international arbitration *per se*. Instead, these disputes are generally analyzed as a matter of trade law.

Categories of Analysis

Domestic forms of arbitration and litigation have long been subject to analyses based on law and

economics (Benson 2000; Sanchirico 2012). Although some of these studies may be readily transferable to the international setting, caution must be exercised, given the numerous distinctions between international and domestic forms of dispute resolution.

At one time, scholars hesitated to use law and economics to consider international legal concerns (Dunoff and Trachtman 1999). However, the last 15 years has seen a significant increase in economic analyses relating to international law. Some of these studies have focused on institutional concerns, while others have emphasized features relating to the allocation of jurisdictional authority in matters involving regulatory overlap (Guzman 2008a; Trachtman 2008; Danielsen 2011). A significant amount of work has also been done in matters concerning international litigation and arbitration, facilitated, in large part, by the ever-increasing amount of empirical data concerning international dispute resolution (Drahozal 2006; Franck 2007a; Van Harten 2012).

At this point, the law and economics literature on international litigation and arbitration appears to fall into four different categories: effectiveness of international dispute resolution, competition between and within different dispute resolution mechanisms, choice of substantive and procedural law and conflict of laws, and third-party funding for international litigation and arbitration. Each of these matters is addressed separately below.

Effectiveness of International Dispute Resolution

Domestic forms of dispute resolution are typically predicated on assumptions about the coercive power of the state. However, some forms of international dispute resolution involve states as defendants or respondents, thus raising questions about whether, why, and to what extent such mechanisms are effective, given the absence of traditional means of coercion. These issues have been analyzed from several perspectives, including that of rational choice theory (Guzman 2008b), game theory (Ginsburg and McAdams 2004), and behavioral economics (Van Aaken 2014).

The premises of the individual studies can vary significantly. In some cases, researchers evaluate the effectiveness of different international tribunals so as to help improve the design of future international dispute systems and thereby maximize the possibility that the legal system in question will achieve its stated goals. Other scholars study effectiveness in order to identify the limitations of the relevant court or tribunal. In all cases, the ultimate aim is to rationalize a field that has traditionally existed outside the realm of empirical and scientific analysis. When considering these issues, commentators have discussed bodies as diverse as the International Court of Justice, the European Court of Human Rights, the International Tribunal for the Law of the Sea, and various sorts of arbitral tribunals, including those operating under the Convention on the Settlement of Investment Disputes Between States and Nationals of Other States (ICSID Convention).

Competition Between and Within Different Dispute Resolution Mechanisms

International dispute resolution has become increasingly fragmented in the last few decades. Although international courts and tribunals may have exclusive jurisdiction over some types of legal injuries, parties can often bring the same or similar claim in another venue. As a result, questions regarding jurisdiction can be framed in terms of both commons and anticommons.

The existence of jurisdictional choice has resulted in a great deal of competition between and within different dispute resolution mechanisms, which has led to scholars in law and economics to consider these sorts of issues from a variety of perspectives (Ramsmeyer 2000; Trachtman 2008). Perhaps the most extensive work has been done on the relative benefits of international commercial arbitration over international litigation in national courts. One natural area of inquiry involves the question of transaction costs, which in this field can include costs associated with the dispute resolution process itself (often referred to as "litigation costs") as well as costs associated with negotiation of the underlying dispute resolution provision.

Most scholarship in this field has focused on litigation costs, which makes sense in light of the traditional understanding of international commercial arbitration as a less expensive option than international litigation in national courts. However, the cost-effectiveness of arbitration has recently been called into question, at least with respect to direct costs such as those relating to attorneys' fees and arbitrators' fees. As a result, some observers have suggested that arbitration has lost its competitive advantage over litigation.

Although the increase in direct costs is disturbing to many in the international community, international commercial arbitration may nevertheless retain its superiority over litigation because arbitration allows parties to reap certain indirect benefits. For example, international commercial arbitration is often said to reduce transaction costs and provide a more efficient means of resolving international commercial disputes because it offers more predictability with respect to the forum, procedure, substantive law, and enforceability of the resulting decision (Ramsmeyer 2000; Strong 2014). Although widespread adoption of the Hague Convention on Choice of Court Agreements could apply some pressure with respect to these features, that instrument has not yet come into force. As a result, international litigation does not seem to be closing the gap with arbitration in terms of procedural or substantive predictability.

To the contrary, studies appear to suggest the continuing existence of a "home court advantage" in national courts (Bhattacharya et al. 2007). Furthermore, the law and practice relating to the recognition and enforcement of foreign judgments around the world remains highly unpredictable (Rotem 2010). As a result, the most rational course of action for international commercial actors is to avoid international litigation in either party's national court.

Although this data suggests that international commercial arbitration is preferable to international litigation, law and economics scholars have nevertheless considered whether particular arbitral practices can be improved upon. One area of interest involves the pre-hearing exchange of documents and information (referred to as either disclosure or discovery). Recent years have seen an increase in the amount of information that parties have sought on a pre-hearing basis. However, economic analysis has suggested a rational actor should prefer the more limited form of document production that characterized early forms of international commercial arbitration rather than the more expansive type of disclosure that is increasingly becoming the norm (Rojas Elgueta 2011).

Some commentators have analyzed arbitral procedures on a more comprehensive basis and have factored in whether and to what extent arbitration is capable of providing ancillary assistance (such as the ability to issue anti-suit injunctions or freeze assets) and increasing the enforceability of the resulting decision (Rutledge 2012). These studies suggest that, rather than adapting to make arbitral procedures more like judicial procedures, international commercial arbitration should retain its distinctiveness so as to retain its competitive advantage. Indeed, some scholars believe the comparative benefits of international commercial arbitration are so significant that arbitration should be made the default position in international commercial disputes (Cuniberti 2009). Some commentators distinguish between the efficiency of international commercial arbitration (Posner 1999; Kovacs 2012) and that of investment arbitration (Franck 2011).

One issue that the literature is addressing with increasing frequency involves the role that dispute resolution mechanisms play in the development of international regulatory law. In many cases, the absence of a single political actor with international regulatory authority has required parties who have suffered multijurisdictional legal injuries to seek relief from national courts. Although some judges have occasionally been willing to fill these sorts of international regulatory gaps, national courts are typically limited in their ability to exercise jurisdiction over foreign parties or apply domestic law extraterritorially (Buxbaum 2006). Furthermore, it can be difficult, if not impossible, for national courts to coordinate their regulatory standards and mechanisms, thereby resulting in market inefficiencies.

Scholarship in this field offers a number of different insights and conclusions. For example, some observers suggest that national courts should be allowed to exert a broad jurisdiction over international regulatory concerns so as to increase efficiency in this area (Mehra 2004). This approach is supported by studies suggesting that private litigants may be particularly efficient enforcers of public regulatory aims in cases involving public goods or commons constellations (Van Aaken 2010). However, other commentators have expressed concern about overgeneralization in this area of law and have suggested that the varying nature of different public goods requires case-by-case analysis (Symposium 2012).

One way to avoid some of the problems associated with regulation via litigation in national courts is through arbitration, since arbitration does not experience the same kinds of problems as litigation does with respect to jurisdiction, enforcement, and application of foreign law (Strong 2013). However, it is unclear whether and to what extent it is appropriate for private tribunals to adjudicate matters of public concern. Some commentators have suggested that questions of public law should be arbitrable in developing but not developed countries, since that approach maximizes economic development and international commercial activity (McConnaughay 2001).

Other scholars evaluating regulatory concerns have focused not on the nature of the claim or the quality of the governing law, but instead on the parties themselves. For example, questions have arisen as to whether state actors may be considered proper participants in international commercial (as opposed to international investment) arbitration. Authors approaching the issue from a law and economics perspective have answered that question positively, based on analyses demonstrating that states operate as firm-like economic organizations (Guevera-Bernal 2004).

Although states sometimes appear as parties in international commercial arbitration, they may also be named as respondents in international investment arbitration. Investment arbitration is a risky and costly endeavor, and states and commentators are continually reassessing the economic value of participating in a treaty regime that exposes them to such a high degree of financial liability (Franck 2007b; Trakman 2012; United Kingdom Department for Business, Innovation and Skills 2013). One of the issues that is constantly under scrutiny is whether and to what extent participation in the investment treaty regime actually increases foreign direct investment in the signatory state (Franck 2007b).

States are not the only parties that must evaluate the value of investment arbitration. Investors who believe that they have suffered a legal injury typically undertake sophisticated cost-benefit analyses to determine whether to pursue a claim in investment arbitration or in another possible forum (Bjorklund 2007). Empirical and economic studies help investors make these decisions by illuminating the likelihood of success in a particular venue (Franck 2007a) and the costs associated with seeking a certain type of legal redress (Franck 2011).

Competition not only exists between different types of international dispute resolution, it also exists within each category of procedures. For example, arbitral institutions from around the world are constantly working to optimize their attractiveness to prospective parties and distinguishing themselves from competitor organizations (Dezalay and Garth 1995; Drahozal 2000a; Ramsmeyer 2000; Choi 2003; Rogers 2006). In so doing, arbitral institutions must identify those procedural trends that should be followed and those procedural innovations that should be resisted. Arbitral institutions are now able to differentiate themselves in a variety of ways, ranging from transparency of proceedings to the availability of expedited procedures. Economic analyses consider not only whether and to what extent individual innovations improve the efficiency of the arbitral process but also how external market forces drive competition between the different organizations.

Competition also exists between individual cities and countries that want to position themselves as potential seats for arbitration. Rent-seeking lawyers may undertake a variety of efforts to position their city or state favorably vis-à-vis any actual or perceived competitors (Drahozal 2000a,

2004; O'Connor and Rutledge 2014). Economic analyses in this area consider the effectiveness of these various means of attracting arbitration business to a particular region, including standardization versus innovation in the national arbitration law, improvement of judicial procedures relating to arbitration, liberalization of the underlying substantive law, and relaxation of rules regarding the unauthorized practice of law.

Competition also exists with respect to individual arbitrators seeking to be named to various public or private tribunals (Dezalay and Garth 1995; Drahozal 2000a; Rogers 2005, 2006; Franck 2009). Commentators have suggested that informational imperfections regarding the qualifications and availability of potential arbitrators and the performance of arbitrators create numerous inefficiencies in the process of selecting arbitrators. Furthermore, the market for international arbitrators appears to reflect a number of barriers to entry. Indeed, empirical research clearly demonstrates that the market for international arbitrators is extremely constrained in terms of age, nationality, race, gender, education, and professional qualifications.

Choice of Law and Conflict of Laws

Another area of inquiry that has benefitted from the application of economic analysis involves choice of law in international disputes. A number of studies focus on issues of substantive law, based on the premise that substantive inefficiencies have a detrimental effect on global economic welfare (Drahozal 2000a; Rühl 2006; Whytock 2009; Ller 2011; O'Connor and Rutledge 2014). When testing this hypothesis, some researchers focus on the relative benefits of different national laws and non-state law, such as the International Institute for the Unification of Private Law (UNIDROIT) Principles of International Commercial Contracts and the United Nations Commission on International Trade Law (UNCITRAL) Convention on the International Sale of Goods (CISG), while other commentators consider the transactional costs associated with creating individualized contracts relating to choice of law. Other analyses look at state-imposed conflict of laws rules to determine whether and to what extent these default regimes can be considered either efficient or reflective of the parties' presumed intent. In each of these cases, the methodological approach and motivating principles are similar to those seen in studies of the domestic law market, although the analysis is much more challenging at the international level, since the relatively high degree of substantive and procedural autonomy associated with international dispute resolution increases the number of variables that must be considered.

One area of particular interest to law and economics scholars is the *lex mercatoria*, also known as the international law merchant. The *lex mercatoria* is said to embody certain customary international legal norms concerning trade practices and usages and therefore provides various insights into the nature of contract law and corporations (Symposium 2004; Oman 2005). Other studies have focused on whether and to what extent arbitrators actually rely on the *lex mercatoria* and have often concluded that this body of law is applied much less frequently in practice than in theory (Drahozal 2000a; Symposium 2004).

Although most studies relating to substantive choice of law are global in nature, some focus on the particular pressures that arise within the European Union as a result of regional integration of certain political and economic matters (Watt 2003; Michaels and Jansen 2006). Specialists in European law have also considered the extent to which US-style law and economics theory has really taken hold in European legal thinking regarding conflict of laws questions and have concluded, contrary to certain conventional wisdom, that there is indeed evidence of a cross-Atlantic jurisprudential dialogue (Nicola 2008).

Although the literature on law and economics is rife with discussions relating to choice of substantive law, procedural choice of law is also an issue of interest in the field. Thus, a number of studies consider the transaction costs associated with customizing arbitration provisions (Choi 2003) or litigation procedures (Drahozal and O'Connor 2014; Hoffman 2014; Strong 2014). Although parties often have a great deal of scope in which to exercise their procedural autonomy,

several of these studies suggest that parties seldom do so, preferring instead to rely on default provisions established as a matter of state law or arbitral custom.

Other inquiries focus more on systemic issues. Thus, some commentators consider the economic benefits associated with procedural harmonization across national boundaries (Visscher 2012), while other observers focus on questions relating to why procedures promoting settlement are present in some jurisdictions but absent in others (Cortés 2013). These types of analyses may be of particular relevance to legislators seeking to improve the design of their national dispute resolution regimes.

Legislators may also be interested in studies focusing on the efficacy of specific rules of procedure. Thus, economic analyses have been applied to questions involving the exercise of extraterritorial jurisdiction (Trachtman 2001, 2008), enforcement of foreign judgments (Rotem 2010; Rühl 2006), and enforcement of foreign arbitral awards (Drahozal 2000b). These studies are quite diverse in nature and consider everything from the role of externalities in the decision-making process and how international jurisdiction is both a commons and an anticommons to the role of informational asymmetries and general issues relating to regulatory competition. However, some critics contend that law and economics is not the proper lens through which to view problems of international procedure (Kessedjian 2005).

The literature also encompasses a number of macro-level analyses, such as those considering the differences, if any, between hard and soft forms of international law (Shaffer and Pollack 2010) and between the common law and civil law traditions (Mattei 1997; Parisi 2002; Arruñada and Andonova 2008). These studies can be useful in determining the cost-effectiveness of various types of reform, both as a substantive and procedural matter.

Third-Party Funding

As international dispute resolution has become more expensive, parties and legal systems have begun to explore alternate ways of funding the various proceedings. One particularly intriguing mechanism involves third-party funding, which is sometimes referred to as "third-party litigation funding." This device involves an unaffiliated entity choosing to pay for the legal costs of one of the parties in exchange for a certain percentage of the final award. Third-party funding has been used in both litigation and arbitration and is coming under increased economic scrutiny. Some studies are comparative in nature (Barker 2012), whereas others are either national (Hylton 2012) or international (Steinitz 2011) in scope. Few conclusions have yet been reached, since the mechanism is still in its early stages, but early analyses suggest that third-party funding will fundamentally change the economics of dispute resolution.

Cross-References

- ▶ Alternative Dispute Resolution
- ▶ Conflict of Laws
- ▶ Globalization
- ▶ Lex Mercatoria

References

General

Arruñada B, Andonova V (2008) Common law and civil law as pro-market adaptations. Wash Univ J Law Policy 26:81–130

Bjorklund AK (2007) Private rights and public international law: why competition among international economic law tribunals is not working. Hastings Law J 59:241–307

Born G (2012) A new generation of international adjudication. Duke Law J 61:775–879

Danielsen D (2011) Economic approaches to global regulation: expanding the international law and economics paradigm. J Int Bus Law 10:23–89

Dezalay Y, Garth B (1995) Merchants of law as moral entrepreneurs: constructing international justice from the competition for transnational business disputes. Law Soc Rev 29:27–62

Dunoff JL, Trachtman JP (1999) Economic analysis of international law. Yale J Int Law 24:1–59

Ginsburg T, McAdams RH (2004) Adjudicating in anarchy: an expressive theory of international dispute resolution. William Mary Law Rev 45:1229–1331

Guzman A (2008a) How international law works: a rational choice theory. Oxford University Press, Oxford

Guzman A (2008b) International tribunals: a rational choice analysis. Univ Pa Law Rev 157:171–235

Guzman A, Simmons BA (2002) To settle or empanel? An empirical analysis of litigation and settlement at the World Trade Organization. J Legal Stud 31:205–227
Kessedjian C (2005) Dispute resolution in a complex international society. Melb Univ Law Rev 29:765–808
Ler HE (2011) The transnational law market, regulatory competition, and transnational corporations. Indiana J Glob Legal Stud 18:707–749
Mattei U (1997) Comparative law and economics. University of Michigan Press, Ann Arbor
McConnaughay PJ (2001) The scope of autonomy in international contracts and its relation to economic regulation and development. Columbia J Transnatl Law 39:595–656
Michaels R, Jansen N (2006) Private law beyond the state? Europeanization, globalization, privatization. Am J Comp Law 54:843–890
Oman N (2005) Corporations and autonomy theories of contract: a critique of the new *lex mercatoria*. Denver Univ Law Rev 83:101–145
Rutledge PB (2012) Convergence and divergence in international dispute resolution. J Dispute Resol 2012:49–61
Rühl G (2006) Methods and approaches in choice of law: an economic perspective. Berkeley J Int Law 24:801–841
Sanchirico CW (2012) 8 Encyclopedia of law and economics: procedural law and economics. Edward Elgar, Cheltenham
Shaffer GC, Pollack MA (2010) Hard vs. soft law: alternatives, complements, and antagonists in international governance. Minn Law Rev 94:706–798
Spain A (2010) Integration matters: rethinking the architecture of international dispute resolution. Univ Pa J Int L23:1–55
Symposium (2004) The empirical and theoretical underpinnings of the law merchant. Chic J Int Law 5:1–190
Symposium (2012) Global public goods and the plurality of legal orders. Eur J Int Law 23(3):643–791
Trachtman J (2008) The economic structure of international law. Harvard University Press, Cambridge, MA
Van Aaken A (2010) Trust, verify, or incentivize? Effectuating public international law regulating public goods through market mechanisms. Am Soc Int Law Proc 104:153–156
Van Aaken A (2014) Behavioral international law and economics. Harv J Int Law 55
Watt HM (2003) Choice of law in integrated and interconnected markets: a matter of political economy. Columbia J Eur Law 9:383–409
Whytock C (2009) Myth of mess? International choice of law in action. NY Univ Law Rev 84:719–790

International Arbitration

Benson BL (2000) Arbitration. In: Bouckaert B, De Geest G (eds) Encyclopedia of law and economics: the economics of crime and litigation. Edward Elgar, Cheltenham, pp 159–193
Choi SJ (2003) The problem with arbitration agreements. Vanderbilt J Transnatl Law 36:1233–1240
Cuniberti G (2009) Beyond contract – the case for default arbitration in international commercial disputes. Fordham Int'Law J 32:417–488
Drahozal CR (2000a) Commercial norms, commercial codes, and international commercial arbitration. Vanderbilt J Transnat Law 33:79–146
Drahozal CR (2000b) Enforcing vacated international arbitration awards: an economic approach. Am Rev Int Arb 11:451–479
Drahozal CR (2004) Regulatory competition and the location of international arbitration proceedings. Int Rev Law Econ 24:371–383
Drahozal CR (2006) Arbitration by the numbers: the state of empirical research on international commercial arbitration. Arb Int 22:291–307
Franck SD (2007a) Empirically evaluating claims about investment treaty arbitration. N C Law Rev 86:1–87
Franck SD (2007b) Foreign direct investment, investment treaty arbitration and the rule of law. McGeorge Glob Bus Develop Law J 19:337–373
Franck SD (2009) Development and outcomes of investment treaty arbitration. Harv Int Law J 50: 435–489
Franck SD (2011) Rationalizing costs in investment treaty arbitration. Wash Univ Law Rev 88:769–852
Guevera-Bernal I (2004) The validity of state contracts arbitration: a 'law and economics' perspective. Revista de la Maestría en Derecho Económico 2:7–20
Kovacs RB (2012) Efficiency in international arbitration: an economic approach. Am Rev Int Arb 23: 155–174
O'Connor EO, Rutledge P (2014) Arbitration, the law market, and the new law of lawyering. Int Rev Law Econ 38:87–106
Posner E (1999) Arbitration and the harmonization of international commercial law: a defense of *Mitsubishi*. Va J Int Law 39:647–670
Ramsmeyer JM (2000) International dispute resolution: Law and economics. In: Hamada K et al (eds) Dreams and dilemmas: economic friction and dispute resolution in the Asia-Pacific. Institute of Southeast Asian Studies, Singapore, pp 464–477
Rogers CA (2005) The vocation of the international arbitrator. Am Univ Int Law Rev 20:957–1020
Rogers CA (2006) Transparency in international commercial arbitration. Univ Kans Law Rev 54: 1301–1337
Rojas Elgueta G (2011) Understanding discovery in international commercial arbitration through behavioral law and economics: a journey inside the minds of parties and arbitrators. Harv Neg Law Rev 16:165–191
Strong SI (2009) Research and practice in international commercial arbitration: sources and strategies. Oxford University Press, Oxford
Strong SI (2013) Mass procedures as a form of "regulatory arbitration" – *Abaclat v. Argentine Republic* and the

international investment regime. J Corp Law 38:259–324
Trakman LE (2012) The ICSID under siege. Cornell Int Law J 45:603–665
United Kingdom Department for Business, Innovation & Skills, Analytical Framework for Assessing Costs and Benefits of Investment Protection Treaties (2013) Available at https://www.gov.uk/government/publications/analytical-framework-for-assessment-costs-and-benefits-of-investment-protection-treaties
Van Harten G (2012) Arbitrator behaviour in asymmetrical adjudication: an empirical study of investment treaty arbitration. Osgoode Hall Law J 50:211–268
Welsh NA, Schneider AK (2013) The thoughtful integration of mediation into bilateral investment treaty arbitration. Harv Neg Law Rev 18:71–144

International Litigation

Barker GR (2012) Third-party litigation funding in Australia and Europe. J Law Econ Policy 8:451–524
Bhattacharya U et al (2007) The home court advantage in international corporate litigation. J Law Econ 50:625–659
Buxbaum HL (2006) Transnational regulatory litigation. Va J Int Law 46:251–317
Cortés P (2013) A comparative review of offers to settle – would an emerging settlement culture pave the way for their adoption in continental Europe? Civil Justice Quart 23:42–67
Drahozal CR, O'Connor EO (2014) Unbundling procedure. Fla L Rev 66:389–430
Hoffman DA (2014) Whither bespoke procedure? Ill Law Rev
Hylton KN (2012) The economics of third-party financed litigation. J Law Econ Policy 8:701–741
Mehra SK (2004) More is less: a law-and-economics approach to the international scope of private antitrust enforcement. Templ Law Rev 77:47–70
Nicola FG (2008) Transatlanticisms: constitutional asymmetry and selective reception of U.S. law and economics in the formation of European private law. Cardozo J Int Comp Law 16:87–153
Parisi F (2002) Rent-seeking through litigation: adversarial and inquisitorial systems compared. Int Rev Law Econ 22:193–216
Rotem Y (2010) The problem of selective or sporadic recognition: a new economic rationale for the law of foreign country judgments. Chic J Int Law 10: 505–537
Steinitz M (2011) Whose claim is this anyway? Third-party litigation funding. Minn Law Rev 95:1268–1338
Strong SI (2014) Limits of procedural choice of law. Brooklyn J Int Law 39
Trachtman J (2001) Economic analysis of prescriptive jurisdiction. Va J Int Law 42:1–79
Visscher L (2012) A law and economics view on harmonisation of procedural laws. In: Kramer XE, Rhee CH (eds) Civil litigation in a globalising world. Springer, New York, pp 65–92

Internet Governance

Philip C. Hanke
Department of Public Law, University of Bern, Bern, Switzerland

Abstract

The Internet is not regulated by a central authority, but instead by a plethora of institutions involving numerous stakeholders. Internet governance can be understood as a field that not only pertains to infrastructure regulation, but also to other legal questions such as copyright, hate speech, cyber-bulling, data protection, and others.

Definition

Internet governance, according to the UN-initiated World Summit on the Information Society, is "the development and application by Governments, the private sector and civil society, in their respective roles, of shared principles, norms, rules, decision-making procedures, and programmes that shape the evolution and use of the Internet" ("Report of the Working Group on Internet Governance" 2005).

Introduction

The Internet is a globally distributed network constituting a patchwork of interconnected autonomous networks. As such, there is no centralized governing body overseeing it. However, a governance structure is necessary for coordination and the assignment of certain property rights (e.g., domain names and Internet Protocol addresses). One early characterization of Internet governance identified three layers of regulation: the physical infrastructure layer contains the hardware, the logical layer contains the software, and the content layer contains the transmitted information. Each layer comes with its own questions, such as network access, copyright issues, or monetization (Benkler 2000).

The Internet, in a way, can be considered a public good (Spar 1999). Consumption is basically non-rivalrous. Although congestion can be an issue with increasing usage, its infrastructure is highly scalable and expandable almost without limits. The Internet also fulfills the second requirement of public goods, namely, that users cannot be excluded from accessing it. Certain services (e.g., access through specific Internet providers) are excludable, but the architecture as a whole is not.

Despite its characteristic as a public good, it has nevertheless successfully transferred into the private sector over the last 20 years. Although it first developed at universities and government agencies, the infrastructure, and arguable most of its content, is provided by private companies.

A Brief History of Internet Governance

The Internet as it is known today evolved from several precursor networks. One important network was ARPANET, developed and funded by the Advanced Research Projects Agency (ARPA) of the US Department of Defense (DoD) and established in 1969. In 1981, it was expanded by the National Science Foundation's Computer Science Network to connect its supercomputing facilities, research networks, and university campus networks. Since 1982, ARPANET has used the TCP/IP protocol suite, which is the cornerstone of modern Internet infrastructure. Throughout the 1980s, other countries were connected to ARPANET/NSFNET. In the 1990s, the network became the Internet: ARPANET was terminated in 1990, and the NSF started to first allow commercial usages of NSFNET in 1991, lifting the final restrictions in 1995. The NSF ended its sponsorship of the backbone services, and the private sector took over the provision of infrastructure. An originally state-owned network had thus become a mainly private network regulated by standards and commercial agreements (for a historical overview, see, e.g., Abbate 2000; Ryan 2013). Certain legacies from the Internet's early era, such as the role of the US government, survived until very recently.

Who Governs the Internet?

The operation of the Internet requires certain administrative tasks, such as assigning domain names (e.g., google.com), IP addresses (e.g., 216.58.214.110 is an IP address that the domain name google.com points to – however, a large website like Google operates many servers around the globe and thus uses many IP addresses for their different services), and other parameters. This assignment of names and numbers is overseen by the nonprofit Internet Corporation for Assigned Names and Numbers (ICANN) based in Los Angeles, California. ICANN, in turn, is managed by an International Board of Directors representing the constituencies of ICANN, its Supporting Organizations (the Generic Names Supporting Organization, the Country Code Names Supporting Organization, and the Address Supporting Organization), and its subgroups. Governments enter the structure through the Governmental Advisory Committee, which has an advisory role.

The Internet's domain name system (i.e., the list that assigns domain names to IP numbers) is administered by the Internet Assigned Numbers Authority (IANA), managed by ICANN. Until 2016, the US Department of Commerce (DoC) could overrule any resolution made by that body. This special role of the US government was a contentious issue for a long time, and handing over full control over IANA to ICANN was an important step in fully transition Internet governance toward a fully multilateral, international system.

An important forum to develop Internet governance is the Internet Governance Forum (IGF). It was established in 2006 by the secretary-general of the United Nations and holds annual meetings. Its main institutional bodies are the Multistakeholder Advisory Group (MAG) and the Secretariat based in Geneva, Switzerland. The MAG currently consists of 55 members representing governments, private industry, civil society, academia, and technical communities. Membership is not permanent but instead rotates regularly, and the UN secretary-general has the final say in the selection process initiated by the undersecretary-general for Economic and Social Affairs.

The standardization of the Internet's core protocols takes place under the auspices of the Internet Society (an NGO founded in 1992 and based in Reston, Virginia, and Geneva, Switzerland), more specifically mainly within the Internet Engineering Task Force, an open standards organization run by volunteers (whose work is usually funded by their employers). It also serves the host institution for the Internet Architecture Board (a committee of the IETF), the Internet Engineering Steering Group (a committee responsible for the technical management of IETF activities), and the Internet Research Task Force (focused on longer-term research issues related to the Internet).

Standards relating to the World Wide Web, that is, the part of the Internet that is accessed through the hypertext transfer protocol (HTTP) using a web browser, are set by World Wide Web Consortium (W3C). The Consortium has well over 400 members, including NGOs, private businesses, academic institutions, governmental entities, and some private individuals. Membership is subject to fulfilling certain requirements and receiving the approval of the existing members.

Thus, Internet governance is a patchwork, and proposals for reform to construct a coherent regime have been made since the mid-1990s. A global framework, for example, through a convention, was discussed in the academic debate but did not find momentum at the policy level (see, e.g., Mueller et al. 2007; Weber 2014).

Contentious Issues

The view that the Internet is a border-free network, free from territoriality, and thus controlled by governments is, however, contested. It has been argued that in many Internet-related policy fields – such as e-commerce, privacy, speech and pornography, intellectual property, and cybercrime – national governments do play a dominant role in regulation (Goldsmith and Wu 2006).

Occasionally, big decisions regarding the infrastructure arise. One large issue was a change in the length of Internet Protocol (IP) addresses, that is, the unique numbers required to identify machines on the Internet to be able to exchange information. The original system (IPv4) allowed for 4.3 billion addresses, which is not enough for modern uses of the Internet. Although a new standard – IPv6, allowing for 340 undecillion addresses (36 zeros) – was selected already in the 1990s, the conversion process is still ongoing, which has to do with a nonobvious yet present political process taking place in the background (DeNardis 2009). The interconnectedness between infrastructure standard setting and broader, more political questions such as openness and transparency can thus make fairly straightforward decisions like increasing the number of IP addresses more difficult.

Another highly contested issue in terms of infrastructure is net neutrality, that is, the principle that Internet service providers (ISPs) and regulatory agency should treat all data on the Internet the same – the Internet equivalent of the common carrier principle in common law countries and the public carrier principle in civil law countries. In the European Union, Art. 3(3) of EU Regulation 2015/2120 requires ISPs to "treat all traffic equally, when providing Internet access services, without discrimination, restriction or interference, and irrespective of the sender and receiver, the content accessed or distributed, the applications or services used or provided, or the terminal equipment used." Some EU member states also have national laws ensuring net neutrality. In the United States, net neutrality has been the object of many years of lobbying. In 2015, the Federal Communications Commission set forth rules protecting the equal treatment of Internet traffic, mainly by classifying broadband as a common carrier. However, it is possible that these rules change again under the Trump administration. The policy of mandated net neutrality is highly controversial: while it is defended on the grounds that it protects consumers by keeping access to the Internet equal to all and is important for innovation (Lee and Wu 2009), some economists criticize and see it as price regulation hindering competition (Hahn and Wallsten 2006; Becker et al. 2010; see in particular Maillé et al. 2012 for an overview of the debate).

Conclusion

A narrow definition of Internet governance would solely address questions of Internet infrastructure.

However, a broader view would encompass aspects of policing contents delivered through the Internet and thus include problems such as cyberbullying, copyright law, data protection rules, as well as safety-related matters.

Traditional approaches based on the principles of state sovereignty and clearly defined territoriality would fail to address the inherent problem of providing and maintaining a decentrally provided public good. Besides questions of infrastructure, activities such as hate speech or copyright infringement lead to strong interjurisdictional externalities. As a result, Internet governance is an archetypical example of highly decentralized, multilayered, transnational, public-private governance. Many important institutions for standard setting (such as the IETF) remain rather informal organizations and show a certain degree of ad hocism. Functionally, there is a lot of overlap between infrastructure design, public policy, and education in the activities of the involved institutions. With increasing digitalization and reliance on the Internet, contested questions regarding Internet governance gain relevance as the latter determines Internet freedom and as a result increasingly freedom of society in general (DeNardis 2014).

Cross-References

- ▶ Copyright
- ▶ Hate Groups and Hate Crime
- ▶ Privacy Regulation
- ▶ Telecommunications

References

Abbate J (2000) Inventing the Internet, 58839th edn. The MIT Press, Cambridge, MA

Becker GS, Carlton DW, Sider HS (2010) Net neutrality and consumer welfare. J Competition Law Econ 6(3):497–519. https://doi.org/10.1093/joclec/nhq016

Benkler Y (2000) From consumers to users: shifting the deeper structures of regulation toward sustainable commons and user access. Fed Commun Law J 52(3): 563–580

DeNardis L (2009) Protocol politics. The globalization of internet governance. MIT Press, Cambridge, MA

DeNardis L (2014) The global war for internet governance. Yale University Press, New Haven

Goldsmith J, Wu T (2006) Who controls the internet? Illusions of a borderless world. Oxford University Press, Oxford

Hahn RW, Wallsten S (2006) The economics of net neutrality. Econ Voice 3(6). https://doi.org/10.2202/1553-3832.1194

Lee RS, Wu T (2009) Subsidizing creativity through network design: zero-pricing and net neutrality. J Econ Perspect 23(3):61–76. https://doi.org/10.1257/089533009789176780

Maillé P, Reichl P, Tuffin B (2012) Internet governance and economics of network neutrality. In: Hadjiantonis AM, Stiller B (eds) Telecommunication economics, Lecture notes in computer science, vol 7216. Springer, Berlin, pp 108–116. https://doi.org/10.1007/978-3-642-30382-1_15

Mueller M, Mathiason J, Klein H (2007) The internet and global governance: principles and norms for a new regime. Glob Gov Rev Multilateralism Int Organ 13(2):237–254. https://doi.org/10.5555/ggov.2007.13.2.237

Report of the Working Group on Internet Governance (2005) Château de Bossey. United Nations, Switzerland

Ryan J (2013) A history of the internet and the digital future, Reprint edition. Reaktion Books, London

Spar D (1999) The public face of cyberspace. In: Kaul I, Grunberg I, Stern M (eds) Global public goods: international cooperation in the 21st century. Oxford University Press, Oxford

Weber RH (2014) Realizing a new global cyberspace framework: normative foundations and guiding principles. Springer, Berlin, Heidelberg

Intrinsic and Extrinsic Motivation

Rustam Romaniuc[1] and Cécile Bazart[2,3]
[1]Laboratoire Montpellierain d'économie théorique et appliquée (LAMETA), University of Montpellier; International programme in institutions, economics and law (IEL), University of Turin, Montpellier, France
[2]CEE-M – Univ Montpellier – CNRS – INRA – SupAgro, University of Montpellier, Montpellier, France
[3]Laboratoire Montpellierain d'économie théorique et appliquée (LAMETA), University of Montpellier 1, Montpellier, France

Abstract

The question of human motivation is central to understand the effects of legal rules on people's behavior. One of the main tenets of law and economics is that incentive systems need to be

designed in order to minimize the difference between private and social interests. Legal rules, taxes, subventions, and other external interventions are regarded as necessary to motivate the internalization of externalities. Empirical evidence however suggests that motivation to engage in pro-social behavior may preexist to external incentives. People often avoid cheating, polluting, or littering, and they act pro-socially without considering consequences of deviation to do so. The traditional idea of "laws as price incentives" has a difficult time explaining these phenomena. This is why in the last two decades, one of the most promising research agenda consisted in distinguishing between "extrinsic" and "intrinsic" motivations. The former are linked to actions driven by external incentives, in contrast to the latter driven by personal and internal forces.

Synonyms

External incentives; Internal motivations; Self-interest; Social preferences

Definition

Intrinsic motivation was originally defined as some inherent interest in a task, which cannot be derived from the agent's economic environment. As such, intrinsically motivated behaviors aim at bringing about certain internal rewarding consequences independent of any extrinsic rewards. With the multiplication of experimental results desplaying high levels of cooperation when monetary and social sanctions are inoperative, intrinsic motivation has come to be viewed as a natural response to the idea that socially desirable conduct can solely be motivated by economic incentives. The overarching idea is that because individuals have intrinsic motivations to behave pro-socially, extrinsic incentives such as promises of reward or threats of punishment were not always needed, and, under some conditions, it is desirable to rely on intrinsic motivation alone.

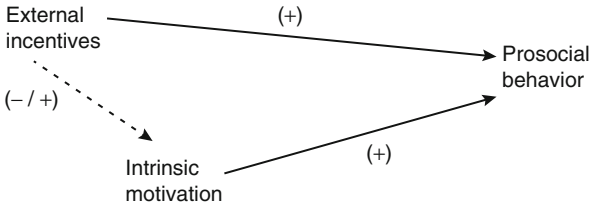
Incentives and Motivations

Selfish decisions, made by one person, can have very bad consequences for other people and society at large. Law and economics refers to this problem as "social costs" (Coase 1960) or "externalities." Thus, it is important to temper individual selfishness to guarantee both: the functioning of the market, which rests on the respect of property rights, and the advancement of social welfare that rides on the provision of public goods. How can we discourage *socially* costly behaviors and encourage people to contribute time and money to the public interest? This question lies at the heart of the "new law and economics" (Kornhauser 1988). This strand of literature suggests that law can create material incentives, such as sanctions or rewards, for even the most selfish individuals to think twice before engaging in socially undesirable conduct. For example, tort law creates such external incentives through the determination of damages the injurer has to pay for the harm caused. This should deter potential injurers from harming others.

However, the limited and sometimes counter effects of external incentives on pro-social behavior are increasingly pointed out by psychologists (Deci and Flaste 1995) and economists (Frey 1992, 1997a, b). And the notion is gaining popularity within the law and economics community (Bohnet et al. 2001; Bohnet and Cooter 2003; Feldman 2009). The idea is that motivation is in fact dual, composed of an intrinsic element besides an extrinsic one.

Intrinsic motivation was originally defined as some inherent interest in a task (Deci 1971). However, it has long since been expanded to include: environmental morale (Frey and Stutzer 2008), civic duty (Frey et al. 1996), fairness and reciprocity (Fehr and Schmidt 1999), and potentially other pro-social unselfish motivations such as "conscience" (Stout 2011).

One of the major insights from this strand of literature is that trying to minimize the difference between individual and social interests by external incentives may actually undermine intrinsic motivation (Gneezy and Rustichini 2000; Bénabou and Tirole 2003). The two types of



Intrinsic and Extrinsic Motivation, Fig. 1 The interactions between external incentives and intrinsic motivation in increasing pro-social behavior

motivations are effectively interacting to produce the final behavioral response. This is called the inseparability thesis (Bowles 2011). Figure 1 illustrates the relationship among these variables, with the positive and negative signs indicating the direction of the effect.

At the core of the interrelationship between intrinsic and extrinsic motivations are the notions of *competence* and *autonomy* (Ryan and Deci 2000; Tirole 2009; Festré and Garrouste 2014). The prescriptive character of rewards and sanctions, such as imposed by the legal system, may indeed undermine people's confidence in their capacities to voluntarily restraint from undesirable conduct (Bénabou and Tirole 2003) and reduce their sense of autonomy by shifting the locus of control from inside to outside the person (Ryan and Deci 2000). Whether this substitution effect exists and to what extent it plays a part in people's reactions to external incentives is a debated empirical question to which we turn next.

Empirical Tests of Motivation Crowding (Out)

Research on the "crowding-out" effects of extrinsic incentives on people's intrinsic motivations can be divided into two groups. The first group focuses on the effects of rewards and sanctions on people's decision to engage in pro-social behavior. The second strand of literature takes other compliance instruments, which are generally labeled as "control" strategies, and attempts to shed light on their capacity to encourage desired conduct.

The Effects of Material Incentives

Contrary to economic theory's postulate that people supply more of a good or service when its price goes up (holding all other prices constant), Titmuss's (1970) empirical investigation of blood donations in the United States of America and Great Britain indicates that paying people in return for their blood might decrease altruistic blood donations. The explanation being that under this monetary incentive, donors no longer consider their act as an altruistic blood donation but as an economic transaction. Following this puzzling idea, Frey et al. (1996) designed a field experiment to test the effects of monetary incentives on Swiss households' acceptance of a nuclear waste recycling plant in their neighborhood. They found that tangible rewards decrease acceptance due to the crowding-out of "public spirit." Along similar lines, Fehr and Rockenbach (2003) suggest that when people attribute their chosen behavior to external sanctions, they discount any altruistic motivations. Curiously, in their game, the amount of cooperation is higher in the absence of the sanction for noncooperative behavior. This is congruent with Bohnet et al. (2001) findings that provisions with respect to contract enforcement may crowd-out trustworthiness and negatively impact performance. In addition, an important aspect of this literature is that external incentives' detrimental effects on cooperation are nonmonotonic. In the case of contract enforcement, performance rates are high when the expected cost of breach is sufficiently large, decreasing for medium-sized expected sanctions and increasing again for sufficiently small ones. With respect to rewards for performance, there is a discontinuity at the zero payment in the effect of monetary incentives: "for all positive but small enough compensations, there is a reduction in performance as compared with the zero compensation, or, better, with the lack of any mention of compensation" (Gneezy and Rustichini 2000, p. 802). Another important

component of the crowding-out theory suggests that people do not really care about the outcome of a pro-social behavior per se, but instead care about how their behavior is perceived by others – they like to be perceived as fair (Andreoni and Bernheim 2009). Their behavior is thus driven by "image motivation" (Ariely et al. 2009). Within this perspective, extrinsic incentives have a detrimental effect on pro-social behavior due to the crowding-out of image motivation. Finally, Irlenbusch and Sliwka (2005) have shown that once a monetary incentive mechanism is introduced, its detrimental effect persists even after its removal and it spreads out to other areas by altering people's cognition of the situation.

The Effects of Control

Another central tenet of economic theory, which is at the core of many law and economics models, is that monitoring improves agents' performance. The legal system often specifies in advance what type of conduct is admissible and the conditions under which it may be carried out. However, the postulate that monitoring is always performance enhancing is jeopardized by experimental results. For instance, Falk and Kosfeld (2006) designed a principal-agent game in which the principal could specify in advance a minimum effort requirement for the agent to put in. Curiously, they found that "the decision to control significantly reduces the agents' willingness to act in the principal's interests" (p. 1611). Agents dislike control because it signals principal's distrust in their capacities to perform a task. Agents "seem to believe that principals who control expect to receive less than those who don't, and (. . .) agents' beliefs correlate positively with their behavior" (p. 1612). This result is not confined to laboratory experiments. In an econometric study of 116 managers in medium-sized Dutch firms, Barkema (1995) showed that increases in personal control by superiors resulted in a significant reduction in the number of hours worked.

Finally, prescribing in detail a particular behavior through ex ante regulations has been found more detrimental than intervening ex post by sanctioning (rewarding) the undesired (desired) conduct. Motivation effects are indeed an important aspect to be considered in the tradeoff between ex ante state control and ex post legal intervention. The legal system is less likely to reduce people's sense of autonomy.

References

Andreoni J, Bernheim BD (2009) Social image and the 50–50 norm: a theoretical and experimental analysis of audience effects. Econometrica 77:1607–1636

Ariely D, Bracha A, Meier S (2009) Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. Am Econ Rev 99:544–555

Barkema HG (1995) Do executives work harder when they are monitored? Kyklos 48:19–42

Bénabou R, Tirole J (2003) Intrinsic and extrinsic motivation. Q J Econ 70:489–520

Bohnet I, Cooter R (2003) Expressive law: framing or equilibrium selection? UC Berkley Public Law Research paper no 138

Bohnet I, Frey BS, Huck S (2001) More order with less law: on contract enforcement, trust, and crowding. Am Polit Sci Rev 95:131–144

Bowles S (2011) Machiavelli's mistake. Mimeo, Santa Fe Institute

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

Deci EL (1971) Effects of externally mediated rewards on intrinsic motivation. J Pers Soc Psychol 18:105–115

Deci EL, Flaste R (1995) Why we do what we do: Understanding self-motivation. Penguin, New York

Falk A, Kosfeld M (2006) The hidden costs of control. Am Econ Rev 96:1611–1630

Fehr E, Rockenbach B (2003) Detrimental effects on sanctions on human altruism. Nature 422:137–140

Fehr E, Schmidt K (1999) A theory of fairness, competition, and cooperation. Q J Econ 114:817–868

Feldman Y (2009) The expressive function of trade secret law: legality, cost, intrinsic motivation, and consensus. J Empir Leg Stud 6:177–212

Festré A, Garrouste P (2014) Theory and evidence in psychology and economics about motivation crowding out: a possible convergence? J Econ Surv. https://doi.org/10.1111/joes.12059

Frey BS (1992) Tertium datur: pricing, regulating and intrinsic motivation. Kyklos 45:161–184

Frey BS (1997a) Not just for the money: an economic theory of personal motivation. Edward Elgar, Cheltenham

Frey BS (1997b) A constitution for knaves crowds out civic virtues. Econ J 107:1043–1053

Frey BS, Stutzer A (2008) Environmental morale and motivation. In: Lewis A (ed) Psychology and economic behavior. Cambridge University Press, Cambridge, pp 406–428

Frey BS, Oberholzer-Gee F, Eichenberger R (1996) Te old lady visits your backyard: a tale of morals and markets. J Polit Econ 104:1297–1313

Gneezy U, Rustichini A (2000) Pay enough or don't pay at all. Q J Econ 115:791–810
Irlenbusch B, Sliwka D (2005) Incentives, decision frames, and motivation crowding-out: An experimental investigation. IZA discussion papers 1879
Kornhauser LA (1988) The new economic analysis of law: legal rules as incentives. In: Mercuro N (ed) Law and economics. Kluwer, Boston, pp 27–55
Ryan RM, Deci EL (2000) Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol 25:54–67
Stout L (2011) Cultivating conscience: How good laws make good people. Princeton University Press, Princeton
Tirole J (2009) Motivation intrinsèque, incitations et normes sociales. Rev Écon 60:577–589
Titmuss RM (1970) The gift relationship. Allen and Unwin, London

Irregular Sector

► Informal Sector