# M

## Mafias

Guglielmo Barone[1] and Gaia Narciso[2]
[1]Bank of Italy and RCEA, Bologna, Italy
[2]Trinity College Dublin, Dublin, Ireland

**Abstract**

Organized crime has far reaching economic, political, and social consequences. This entry presents the basic facts on the economics of mafias, with a special focus on the Sicilian case that may serve as a window into other types of criminal organizations. First, the entry provides a brief sketch of the theoretical modelling of the mafia. Then, it reviews the theoretical and empirical work testing the role of geography in the historical origins of the mafia. Finally, the entry explores the economic impact of the mafia in terms of missed opportunities of development, both in the short and in the long run. Different transmission channels are explored.

### Definition

A secret and criminal organization which is engaged in a number of illicit activities such as racketeering, smuggling, trafficking in narcotics, and money laundering. It has a complex hierarchical organization, and its members are expected to follow a number of internal rules. The mafia originated in Sicily in the second part of the nineteenth century and expanded to the United States through Italian emigration. Nowadays, mafia-type organizations are widespread in many countries, especially in Southern Italy, Russia and East Europe, Latin America, China, and Japan.

### Introduction

Organized crime entails deep economic, political, and social consequences. Its presence is pervasive and threatens the functioning of democratic institutions (Allum and Siebert 2003; Bailey and Godson 2000; Fiorentini and Peltzman 1996). Due to its varied features and the lack of empirical data, very few empirical studies have, until recently, investigated organized crime and its impact on the economy. This emerging literature deals with different issues. Some papers focus on the theoretical framework (Dal Bò et al. 2006; Skaperdas 2001), while empirical studies investigate either the origins of mafia, stressing the role of natural resources and land value as key determinants, or its negative economic consequences. Overall, the mafia is found to have a negative impact on growth and GDP per capita; the underlying mechanism may work through lower government efficiency, lower foreign direct investment, distortions in the allocation of public funds, or unfair markets competition.

This entry presents the evidence on the economics of mafias, with a special focus on the

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

Sicilian case. The Sicilian Mafia is a complex phenomenon that acts, at times, within Italian institutions, at times against them. The Sicilian Mafia serves as a window into other types of criminal organizations, such as the Russian Mafia or the Japanese Yakuza (Maruko 2003), which are rooted in the political and socioeconomic spheres. First, the entry provides a brief sketch of the theoretical modeling of the mafia. Then, it reviews the theoretical and empirical work testing the role of geography in the origins of the mafia. Finally, the entry explores the economic impact of the mafia in terms of missed opportunities of development, both in the short and in the long run.

## Theoretical Models of Mafia and the Analysis of Its Origins

On a theoretical ground, this entry adopts the widespread view according to which the mafia is treated as an industry that produces and sells a number of goods and services, such as private protection services, narcotics, or connections with politics (Gambetta 1993, 2000). Such a view is consistent with the historical origins of the mafia. Under this perspective, the supply of protection services plays a key role. According to a consolidated opinion, the mafia emerged in Sicily after (or around) the unification that took place in 1861. In 1876, Leopoldo Franchetti, a Tuscan intellectual, traveled to Sicily to conduct a private inquiry into the political and administrative conditions of the island. The report was published the following year and represents an original and detailed picture of the state of Sicily at that time (Franchetti 2011). The report is the first document of the issues related to the mafia and its permeation through the Sicilian society. The demand for private protection arose in Sicily for two interrelated motives. First, before Italian unification, a series of anti-feudal laws endorsed the opening up of the market for land, which led to an increase in the fragmentation of land property. Second, following the Italian unification in 1861, a weak protection of property rights and a vacuum of power from the recently constituted State amplified landowners' need for protection against illicit expropriation. In such historical moments, the mafia emerged as an industry offering a number of services the State was not able to offer.

This theory has received robust empirical support. Some authors analyze the relationship between land fragmentation and mafia activity in the nineteenth century. From a theoretical viewpoint, in fact, it can be shown that the rise in the number of landowners (following the end of Feudalism in 1812) increased competition for protection, which ultimately led to an upsurge in mafiosi's profits. The data collected from the parliamentary survey conducted by Damiani in 1881 show that the historical presence of mafia in Sicily was indeed more likely to be found in towns where land was more fragmented (Bandiera 2003). The land fragmentation was not the only determinant of the upsurge in the mafia. The end of feudalism and the demise of the Bourbon Kingdom in the South of Italy were accompanied by a rapid increase in the demand of sulfur, of which Sicily became a major exporter between 1830 and 1850. Sulfur was used as an intermediate input in the industrial and chemical production, which was flourishing in France and England in the nineteenth century. Sulfur mines in Sicily were mainly superficial and did not need sophisticated extraction technology. By the end of the nineteenth century, over 80% of world sulfur production originated from Sicily. Consistently with the idea of mafia as the supply side of a market of protection and extortions, data on the Sicilian Mafia in the late nineteenth century (Cutrera 1900) shows that the intensity of mafia activity was higher in municipalities with sulfur mines, where the demand for private protection was higher Buonanno et al. (forthcoming). In a similar fashion, other authors have associated the origins of the mafia with the presence of citrus fruits, which were highly valuable (Dimico et al. 2012). Lemon trade between Sicily and the United States flourished: 34% of imported citrus fruits originated from Italy. This explanation of the origins of the Sicilian Mafia is indeed in line with that outlined in a parliamentary inquiry dated 1875, according to which "Where wages are low and peasant life is less comfortable, [. . .], there are no symptoms of mafia [. . .]. By contrast, [. . .] where property is divided, where there is plenty of work for everyone, and the orange trees enrich landowners and growers alike – these are the typical sites of mafia influence."

Although the mafia historically emerged in the South of Italy, it gradually migrated to other Italian regions (and to the United States). The first main mechanism of diffusion of the mafia was the mass migration from Southern to Northern Italy, which took place between the 1950s and 1970s. It is estimated that four million people migrated from the South to the North, during the economic boom. The mass migration led to a change in the population composition of the receiving regions. As a result, mafia-type organizations were more likely to emerge in areas that were more migrant-abundant. The second mechanism of the expansion of the mafia in the North is due to the *confino* law: the imprisonment policy for mafiosi was based on the *confino*, a policy according to which mafia-related criminals were imprisoned in a different region from the one they originated from, in order to loosen the links with the local mafia. However, the *confino* had the perverse effect of spreading mafia activity in other Italian regions (Varese 2006, 2011).

## Economic Consequences of Mafia: GDP Growth

Besides analyzing the origins of mafia, economists also examined its consequences in terms of economic growth, together with the potential underlying mechanisms. This is the second main strand of the economic literature on organized crime. Assessing the economic impact of mafia activity suffers from two main issues. The first issue is concerned with the lack of data. Only recently, new data on criminal activities have been made available. However, even where available, data on crime often suffer from measurement error. For example, the number of crimes could be underreported, in particular in relation to specific categories. Second, and more severely, analyzing the impact of mafia activity on the economy implies knowing how the economy would have been in the absence of mafia activity. However, a country or a region is either mafia-ridden or not (this is known as the fundamental problem of causal inference); therefore, it is very difficult to have a credible counterfactual. A convincing way to tackle this issue is to adopt the synthetic control method, a methodology that has been recently proposed to statistically examine comparative case studies. This methodology has originally been introduced to estimate the effect of the Basque conflict on GDP per capita (Abadie and Gardeazabal 2003). Later, it has also been applied to estimate the impact of the mafia in two Southern Italian regions – Apulia and Basilicata – which experienced a surge in mafia activity starting over the last few decades. Until the 1970s, these two Southern regions had witnessed little or no mafia activity in their territory. Starting from the 1970s, mafia activity and its connected violence rapidly increased in these two regions. A counterfactual is constructed using the information related to other regions where the mafia has been absent throughout the period considered. Mafia activity emerged in Apulia and Basilicata following three relevant episodes. First, the profitable activity of tobacco smuggling led to a ferocious conflict among different criminal groups. Second, the earthquake that struck the area between Campania, Basilicata, and Apulia in 1980 was followed by a flow of public funding for the reconstruction of the area. The increased availability of public funding led to an increase in mafia activity, attracted by the opportunities of grabbing a portion of these funds. Finally, the imprisonment policy for mafiosi was based on the *confino*, the policy according to which mafia-related criminals were dislocated in a different region from the one they originated from, in order to loosen their links with the local mafia. Apulia had the highest number of criminals according to the *confino* policy, which led to the spread of mafia activity in Apulia. The impact of mafia organizations on economic growth appears to be very significant. According to the synthetic control estimates, mafias are responsible for a 16% loss in GDP per capita over a 30-year period Pinotti (forthcoming).

## Economic Consequences of Mafia: Misallocation of Public Funds

Such a huge impact captures the reduced-form causal effect from the mafia to GDP. Researchers have also devoted their efforts to highlight the underlying transmission mechanisms. One relevant channel consists of the misallocation of public funds (Barone and Narciso 2015). The case study regards the Italian Law 488/92, which has been the main policy used by the central government to promote growth in the southern Italian regions, by offering a subsidy to businesses investing in underdeveloped areas. Even though the law governing funding allocation had very detailed provisions to distribute subsidies, aimed at reducing the risk of fraud, many investigative reports state that the mafia managed to circumvent these criteria. A number of accounting and financial mechanisms have been used to divert funds, such as the creation of made-up firms, i.e., firms set up with the only scope of applying for public subsidies. Moreover, the mafia was able to corrupt public officials involved in the allocation of funds. By combining information on the spatial distribution of Law 488/92 funds with that of mafia activity, the study investigates whether mafia presence is able to influence public funds' allocation. The endogeneity of the link between the mafia and public funding must be addressed, in order to provide a causal interpretation to their relationship. Endogeneity may arise due to measurement error, omitted variables, and reverse causality. Instrumental variable identification strategy is a proper way to deal with the endogeneity issue. Focusing on Sicily, it is possible to construct a credible instrumental variable that is conceptually based on the historical origins of the Sicilian Mafia. As stated above, around the Italian unification in 1861, the demand for private protection arose in Sicily for two main motives. First, before the country's unification, a series of anti-feudal laws led to the opening up of the market for land, which contributed to the division of land property. Second, the vacuum of power following Italian unification and lack of protection of property rights amplified landowners' need for protection against expropriation. The Sicilian Mafia arose as an industry offering private protection in this specific historical juncture. Indeed, in line with the literature on the origins of the mafia presented above, the mafia was more likely to appear in areas where the land was more valuable. Consequently, historical and geographical measures of land productivity can be used as instrumental variables for current mafia activity. In particular, the set of instruments includes rainfall shocks in the nineteenth century and geographical features (e.g., altitude and slope). This empirical strategy suggests that mafia presence has a positive effect on the likelihood of obtaining funding and the amount of public transfers: according to the existing estimates, mafia presence raises the probability of receiving funding by 64% and increases the amount of public funds to businesses by more than one standard deviation. The results are robust to different econometric specifications and to the use of alternative measures of the mafia. The mafia has a positive causal effect on public subsidies, but how should this result be interpreted? For example, is the positive relationship between mafia presence and public transfers due to a more generous public spending towards mafia-ridden areas? A falsification test suggests that the answer is negative: these areas display a lower level of expenditure on culture and education, in comparison to those where the mafia is absent. Were the State more generous towards disadvantaged areas, presumably it would have spent on other budget categories, such as education or culture. Moreover, there is empirical evidence on the mechanism through which the mafia can determine the allocation of public resources: the mafia is used to make connections with local entrepreneurship, and its presence raises the number of corruption episodes in the public administration sector. The impact of the mafia can also be disentangled from that stemming from a more general criminal environment. In the end, by diverting public subsidies assigned to poorer areas, the mafia hampers growth, investment, and, ultimately, development. From this perspective, this finding provides a relevant contribution to the debate on the desirability and the design of public subsidies to firms.

## Economic Consequences of Mafia: Other Channels

The impact of mafia organizations has been found to also affect the bank credit market. Using data on bank-firm relationships, it has been shown that crime has a negative effect on access to credit in Italy. Borrowers in high-crime areas pay higher interest rates and pledge more collateral. These results are found to be driven, in particular, by mafia-related crime, extortion, and fraud. By distorting loan conditions to firms, mafia organizations indirectly negatively affect investment and growth in the long run.

The effect of mafia on the economy can also work through the public sector. Organized crime usually operates as a pressure group that uses both bribes and the threat of punishment to influence policy. Consistently, on the empirical side, one can observe that criminal activity before elections is correlated with lower human capital of elected politicians and an increased probability that these politicians will be later involved in scandals. However, a strict legal institutional framework can contrast this bad influence. In the Italian case, when the central government imposed the dissolution of municipal government because of mafia infiltration, the quality of local politicians (proxied by their average education level) significantly improved Daniele Geys (forthcoming).

Finally, organized crime hinders fair market competition because mafia-related firms, which are usually run to launder money, can operate under the break-even point, thus forcing legal competitors out of the market; mafia also represents a deterrent for foreign investors, so depressing foreign direct investment.

## References

Abadie A, Gardeazabal J (2003) The economic costs of conflict: a case study of the Basque country. Am Econ Rev 93:113–132

Allum F, Siebert R (2003) Organized crime and the challenge to democracy. Routledge, Abingdon

Bailey J, Godson R (2000) Organized crime and democratic governability: Mexico and the U.S.-Mexican borderlands. University of Pittsburgh Press, Pittsburgh

Bandiera O (2003) Land reform, the market for protection, and the origins of Sicilian mafia: theory and evidence. J Law Econ Organ 19:218–244

Barone G, Narciso G (2015) Organized crime and business subsidies: Where does the money go?. Journal of Urban Economics 86:98–110

Bonaccorsi di Patti E (2009) Weak institutions and credit availability: the impact of crime on bank loans. Bank of Italy occasional papers, no. 52

Buonanno P, Pazzona M (2014) Migrating mafias. Reg Sci Urban Econ Elsevier 44(C):75–81

Buonanno P, Durante R, Prarolo G, Vanin P (forthcoming) Poor institutions, rich mines: resource curse and the origins of the sicilian mafia. Econ J

Cutrera A (1900) La Mafia e i Mafiosi. A Reber, Palermo

Dal Bó E, Dal Bó P, Di Tella R (2006) Plata o pomo: bribe and punishment in a theory of political influence. Am Polit Sci Rev 100:41–53

Daniele G, Geys B (forthcoming) Organized crime, institutions and political quality: empirical evidence from Italian municipalities. Econ J

Dimico A, Isopi A, Olsson O (2012) Origins of the Sicilian mafia: the market for lemons. Working papers in Economics of the University of Gothenburg, no. 532

Fiorentini G, Peltzman S (1996) The economics of organised crime. Cambridge University Press, Cambridge, UK

Franchetti L (2011) Condizioni politiche e amministrative della Sicilia. Donzelli Editore, Roma

Gambetta D (1993) The Sicilian mafia: the business of private protection. Harvard University Press, Cambridge, MA

Gambetta D (2000) Trust: making and breaking cooperative relations, 49–72. Basil Blackwell, New York

Maruko E (2003) Mediated democracy: Yakuza and Japanese political leadership. In: Felia A, Renate S (eds) Organized crime and the challenge to democracy. Routledge, London

Pinotti P (forthcoming), The economic consequences of organized crime: evidence from Southern Italy. Econ J

Skaperdas S (2001) The political economy of organized crime: providing protection when the state does not. Econ Gov 2:173–202

Varese F (2006) How Mafias migrate: the case of the ‘Ndrangheta in northern Italy. Law Soc Rev 40:411–444

Varese F (2011) Mafias on the move: how organized crime conquers new territories. Princeton University Press, Princeton

## Majority Control

► Concentrated Ownership

# Manne, Henry

Enrico Colombatto
Università di Torino, Turin, Italy
IREF (Institut de Recherches Économiques et Fiscales), Lyon, France

**Abstract**

This entry illustrates the very prominent role Henry Manne played in the Law and Economics tradition. Manne made seminal contributions in two key areas: the dynamics of corporate control, and the ethics and efficiency of insider trading. He showed that the market for corporate control is an efficient way of protecting shareholders' interests and restraining abuse by the managers. From a normative standpoint, he emphasized that no specific regulation is required in these areas. The same normative conclusions also apply to insider trading, which is the quickest way of circulating information and avoiding bubbles. This entry concludes by summing up Manne's contributions in education.

## Biography

Henry Manne (1928-2015) deserves a very prominent place among the founding fathers of Law and Economics, along with Guido Calabresi and Ronald Coase. In particular, since the late 1950s, Manne applied economic reasoning to investigate policy issues in two key areas that had previously been considered as the exclusive object of legal analysis: the dynamics of corporate control, and the ethics and efficiency of insider trading. In both cases, and despite considerable initial skepticism, Manne's contributions radically changed the way the economic and legal professions have regarded these topics. Furthermore, and typical of his vision, he used the law-and-economics approach to show that regulating the life of the corporation is unjustified and possibly harmful.

Manne (1962) and Manne (1965) are the path-breaking articles that opened new research agendas in the economics of the modern corporation. In his 1962 contribution, Manne took on the traditional argument according to which small shareholders have little incentive to monitor the managers' behavior and actively partake in the life of companies. In particular, the traditional argument held that regulation is required in order to force the managers to disclose the information shareholders need, to enhance shareholders' participation, and to restrain the managers' potential abusive behavior. By contrast, Manne argued that the market for corporate control is effective in protecting the shareholders' interests, regardless of how much time and efforts they devote to monitoring the managers. If the managers misbehave, the value of the company declines, outsiders will be interested in buying the shares, gaining control and replacing the inefficient executives. In other words, in Mannes' view, the market for mergers and takeovers ensures that share prices cannot drop for long because of managerial slack, as long as outside buyers are allowed to intervene and buy shares to obtain control. Indeed, the threat of a takeover is already enough to prevent the shareholders from being injured: if the incumbent managers fear the consequences of a hostile takeover, they are likely to react by improving their performance and meeting the incumbent shareholders' expectations. If so, performance improves and share prices recover. Once again, shareholders are protected, and share prices stay close to the company's true value. Therefore, Manne concludes, under both circumstances there is no need for regulation/legislation: Competition is indeed effective in keeping the managers under pressure and shielding small investors from abuse. By contrast, regulation frequently ends up defending the managers and justifying redistributive activities that have nothing to do with the purpose of the modern corporation.

Manne (1965) develops his earlier insights by moving from analyzing the relationships among shareholders and managers, to studying the economics of mergers. A takeover takes place when company A buys shares from company B's shareholders. It is an operation that prevents B from going bankrupt and that can be completed in three different ways: proxy fights, direct purchase and mergers. As mentioned, Manne focuses on mergers, the buyers' preferred course of action, since they do not need to collect cash to finance the operation and they can skirt taxation. Not surprisingly, in these cases, corporate statutes play a crucial role in determining the outcome. For example, mergers usually require qualified-majority voting and can hardly succeed if B's incumbent management owns a relatively large portion of B's shares. Thus, one should not be surprised if company A and company B managers end up colluding. Under such circumstances, the merger would then take place with the consent of the management. This is still a desirable outcome: the top executives would be forced to exchange information, which promotes efficiency. As a result, according to Manne the frequent concerns raised by the antitrust authorities are misplaced: mergers are the result of a healthy, free-market, welfare-enhancing environment, rather than a threat to competition or to shareholders' interests.

## Innovative and Original Aspects

Insider trading was the object of Manne (1966), a book that was actually his J.S.D. dissertation thesis at Yale and ignited a debate still lively today. In contrast with the views that dawned at the beginning of the twentieth century and became dominant since the mid-1930s, Manne made two key arguments. First, insider trading is a form of compensation for the managers and employees at large. It is cheaper than other forms of remuneration, and provides sets of incentives that contribute to bringing together the interests of the managers, the entrepreneurs and the owners. Second, insider trading is efficient, since it represents the quickest vehicle to circulate information, and reduces the risk of bubbles. In other words, insiders have accurate information about the company where they work, they ensure that this information is immediately available to the ordinary shareholders, and perform better than the regulators. Hence, legislation designed to curb insider trading is ineffective, harmful, and ethically questionable, since it punishes an alleged crime in the absence of victims.

## Impact and Legacy

Manne's work left a lasting legacy in two other areas: teaching and the economics of higher education. Manne was not only a great scholar, but also a determined and successful intellectual entrepreneur. As detailed in Manne (1993), while serving as a Professor at the University of Rochester, he started planning a new type of law-school curriculum, which emphasized the need for specialization, but at the same time required a cross-disciplinary approach to the selected area of specialization. While still in Rochester, in 1971 Manne launched an intensive summer course in economics designed to suit the needs of law professors. In 1974, Manne moved to the Miami Law School, where the program became the Law and Economics Center at the University of Miami, and offered economics courses to federal judges, too. Later, the Center moved to Emory and then in 1986 to George Mason University, where Manne was heading the Law school and eventually put in practice the innovative curriculum he had drafted in Rochester. These programs were an impressive success: they contributed to the professional life of generations of lawyers, law professors and judges, and made a real difference in the way the legal profession regarded economic issues, in the classroom and in court.

Manne (1973) is perhaps his best-known contribution to the understanding of the nature and dynamics of modern universities. Although focused on the American experience, much of Manne's arguments also apply to the academic environment prevailing in Europe. In particular, Manne drew attention to the consequences of an academic context characterized by extensive funding by governments and governmental agencies, foundations and companies, and in which the role played by the students' fees is modest. This setting has ensured that modern universities are less and less answerable to their clients (students). Put differently, Manne claimed that public funding has ensured that tenured professors are all but unaccountable, and tend to pursue their own interests (mainstream research and consultancy), rather than meet the educational needs of their students, transfer knowledge, and help them develop critical abilities and creative thinking. Moreover, by making students and public opinion trust the virtues of educational establishments as heavily subsidized, not-for-profit organizations, faculties have succeeded in transforming modern universities into stable systems where intellectual entrepreneurship remains marginalized. According to Manne, in order to ensure stability, hiring procedures frequently reward scholarship accomplishment (mainstream research agendas, which do not necessarily imply scholarly value), and favor candidates who guarantee loyalty to the incumbents and feature nonmarket attitudes. The upshot is an ongoing decline in the quality of education, with very few exceptions, on both sides of the Atlantic.

## Cross-References

- ▶ Asymmetric Information in Litigation
- ▶ Economic Analysis of Law
- ▶ Efficient Market
- ▶ Governance
- ▶ Higher Education
- ▶ Law and Economics
- ▶ Law and Economics, History Of
- ▶ Merger Control

## References

Manne H (1962) The 'higher criticism' of the modern corporation. Columbia Law Review, 3, March, pp 399–449

Manne H (1965) Mergers and the market for capital control. J Polit Econ 73(April):110–120

Manne H (1966) Insider trading and the stock market. Free Press, New York

Manne H (1973) The political economy of modern universities. In: Burleigh AH (ed) Education in a free society. Liberty Fund, Indianapolis, pp 165–205

Manne H (1993) An intellectual history of the School of Law George Mason University. the Law and Economics Center, Arlington

## Further Reading

Colombatto E, Cass R. (guest editors) (forthcoming) Symposium in honour of Henry Manne, Eur J Law Econ

McChesney FS (ed) (2009) The collected works of Henry G. Manne., 3 volumes. Liberty Fund, Indianapolis

# Market Definition

Javier Elizalde
Department of Economics, University of Navarra, Pamplona, Navarra, Spain

**Abstract**

The goal of this essay is to explain the role attributed to the analysis of market definition in the guidelines which rule in the United States and the European Union and to revise some of the empirical tests proposed by researchers in the fields of competition Economics and Law along with some of the critiques they received regarding their applicability for antitrust purposes.

## Definition

Market definition is the analysis of determining the products which compete with each other and the geographic area where that competition takes place. The attention is paid, on one hand, to the concept of substitution among potential competitive products and, on the other, to the existence of joint market power by the firms which operate in the market especially for antitrust purposes.

## Introduction

Market definition consists on the delineation of the market boundaries in the product and geographic dimensions from a competitive perspective. The interest for the researchers in the field of Law and Economics has led to the formulation of empirical tests of market definition, and much discussion has arisen mainly regarding their compatibility with the market definition test which rules in the US antitrust legislation for merger control, known as the "hypothetical monopolist" (or SSNIP) test. The analysis of market definition has been performed under two alternative (not necessarily incompatible) purposes. First, in the tradition of Marshall (1920), the market comprises all the goods which are substitutes so they should exhibit identical prices with differences due to transportation costs. This approach is most plausible when the goal of the analysis is to find the geographic market for a homogeneous good. The second approach is to identify the products and the geographic area such that the firms who serve them may jointly enjoy substantial market power without relevant competitive constraints from other products or areas regardless of whether the other alternatives can be considered substitutes from the consumer perspective.

## Market Definition in Merger Control

The market power approach has become predominant for researchers in the field of Law and Economics since the publication of the Horizontal Merger Guidelines in 1982 by the US Department of Justice (see Department of Justice and Federal Trade Commission (2010) for the latest revision). They introduced the hypothetical monopolist test of market definition. The rationale for this test is that the relevant market should include all those products and the geographic area such that, if they were all owned by a single firm (the hypothetical monopolist), the latter would enjoy some market power, being able to profitably exercise a "small but significant and non-transitory increase in price" (the initials SSNIP are used when referring for a price increase with that feature). This means that products other than those forming the relevant market do not impose a significant competitive constraint. The text mentioning the hypothetical monopolist test has been modified in the revisions of the horizontal merger guidelines of 1984, 1992, and 2010, but the role attributed to market definition as a preliminary and screening process in the investigations of competitive effects of mergers prior to the calculation of market shares remains. Market definition is a previous step to the assessment of market power which is the main concern of antitrust authorities when investigating mergers.

In the European Union, there is no test which must be used for market definition. The concept of a relevant market was established in the 1997 "European Commission Notice on the definition of the relevant market for the purposes of Community competition law" and relies on the concept of consumer substitution "[. . .] by reason of the products' characteristics, their prices and their intended use." The definition of a relevant market ruling in the European Union includes both demand-side and supply-side substitutes when substitution takes place "quickly and easily." The Notice mentions a version of the hypothetical monopolist test as a suggested method and enumerates a series of econometric tests, which are actually detailed in the following paragraphs of this essay.

## Empirical Tests of Market Definition

Regarding the empirical literature on market definition, the earliest works were mainly intended to identify the geographic market for a homogeneous good. An approach which was dominant for geographic market definition before the attention was led to prices was the Elzinga-Hogarty test (after Elzinga and Hogarty 1972, 1973) which was a test of shipment data. It used aggregate inflows and outflows of consumers to determine market boundaries. Geographic market boundaries were expanded until both flows were below a cutoff level. This test was used for analysis of hospital mergers in the United States in the 1980s and 1990s, and it received many critiques especially for the limitation of the test to account for the heterogeneity of patients to travel for medical care.

Both in the fashion of the Marshallian concept of a market of equal prices and on the emphasis on the ability to increase prices stated in the US guidelines, the tests of market definition became primarily tests of price data or both prices and quantities when data were available.

There is a vast stream of literature on market definition using tests of time series of prices. Stigler and Sherwin (1985) proposed a test based on price *correlations*. Under the prior that the differences in prices of products of the same market are due to transport costs, their price movements must follow the same pattern so the time series of those prices must be correlated.

Horowitz (1981) suggested that the Marshallian prediction of equality of the prices only occurs in equilibrium. When shocks happen, there are adjustment lags before returning to equilibrium. Horowitz proposed a test called the *speed of adjustment* test as it is focused on estimating the speed at which the short-term difference in prices between two areas returns to the long-term difference. A persistent divergence between the short-term difference and the long-term difference would mean that the two products or areas are in different markets.

Another test called the *causality* test includes products in the same geographic market if the long-run difference between price series is independent of exogenous influences not related to costs. The test, proposed by Uri and Rifkin (1985) and Uri et al. (1985) and based on the concept of Granger causality, defends that the price in one area of a geographic market could be predicted with information on prices in the other area.

Slade (1986) proposed an *exogeneity* test according to which, if a disturbance in one area spills into another area, the exogeneity of price formation is rejected and both areas are in the same market.

A *stationarity* test of market definition was performed by Forni (2004) analyzing the long-run price ratio between two areas rather than a series of short-run price differences which were used in previous tests. Using this test, if the long-run proportional relationship between the prices in two areas is not stationary, then the areas can be said to be in different markets.

On the grounds of market definition related to substitution between products for the consumer, much work has relied on elasticities of demand when data on both prices and quantities has been available. The own-price elasticity of demand provides information on the existence of substitutes for a good, and the cross-price elasticity of demand between two goods provides information on whether those two goods are substitutes, which would be the case if it takes a positive value. A noteworthy approach is the analysis of critical elasticity and critical loss, following the works by Johnson (1989) and Harris and Simons (1989). The main focus of this analysis is the own-price elasticity of demand of the hypothetical monopolist. The critical elasticity is the maximum elasticity of demand a hypothetical profit-maximizing monopolist could face at pre-merger prices to be able to profitably increase its price by a SSNIP. If the elasticity faced at pre-merger prices is higher than the critical elasticity, the hypothetical monopolist would not raise profits with that increase in price so this would lead to market aggregation. The critical loss is the maximum reduction in output a hypothetical monopolist can tolerate in order for the increase in price to be profitable. If the actual loss is higher than the critical loss, the price increase is not profitable, so the relevant market should be expanded.

An alternative approach is based on the own-price elasticity of the residual demand of the hypothetical monopolist. Baker and Bresnahan (1988) used this methodology to estimate the degree of market power in an industry with differentiated products (the beer industry in the United States). Scheffman and Spiller (1987) applied it to the geographic market definition for a homogeneous good (unleaded gasoline in the Eastern United States). Once the value of the own-price elasticity of the hypothetical monopolist's residual demand is estimated, the effect of the price increase on the monopolist's profit is computed (see also Kamerschen (1994), Ekelund et al. (1999), and Cardona et al. (2009) for other works under this approach).

The recent emergence of detailed retail data on prices, quantities, and other variables (often through scanner systems) has favored the estimation of each firm's individual demand, and, through the estimation of both own- and cross-price elasticities of demand, the likely effects of mergers can be simulated. Some works use this methodology for the definition of the relevant market for cars (Brenkers and Verboven 2006, analyzing competition at both the manufacture and retail levels), computer servers (Ivaldi and Lörincz 2011), and movie theaters (Elizalde 2013, which analyzes both demand-side and supply-side substitution).

There is a vast literature, such as Werden (1990, 1998, 2003), Werden and Froeb (1993), O'Brien and Wickelgren (2003), and Hosken and Taylor (2004) among many others, which criticizes most of the tests enumerated above by showing their incompatibility with the hypothetical monopolist test of the US guidelines and their invalidity for predicting the likely effects of mergers, as some of the tests are intended to the identification of substitutes or are based on past evidence with limited capability to predict future events among other reasons.

Coate and Fischer (2008) provide a review of the tests employed by the enforcement agencies in the United States along with comments about the works which proposed them and the critiques they received.

## Controversy Regarding Market Definition

Market definition itself, especially with the role attributed in the US antitrust legislation, is strongly criticized in some grounds of Law and Economics which are very much skeptical about the competition authorities' interest on market concentration and market power, as the immediate usefulness of market definition is the calculation of market shares. Coinciding with the 2010 revision of the US Horizontal Merger Guidelines, an alternative to market definition, based on the upward pricing pressure (UPP) resulting in a merger, was proposed by Farrell and Shapiro (2010) and critically discussed by Carlton (2010), whereas Kaplow (2010, 2011) defends the elimination of market definition as conclusions may be misleading.

## Cross-References

▶ Demand
▶ Economic Analysis of Law
▶ Empirical Analysis
▶ European Community Law
▶ Law and Economics
▶ Merger Control

## References

Baker JB, Bresnahan TF (1988) Estimating the residual demand curve facing a single firm. Int J Ind Organ 6(3): 283–300

Brenkers, R. and F. Verboven (2006): Market definition with differentiated products – lessons from the car market. In: Choi J.P. (ed) Recent developments in antitrust: theory and evidence. MIT Press

Cardona M, Schwarz A, Burcin Yurtoglu B, Zulehner C (2009) Demand estimation and market definition for broadband internet services. J Regul Econ 35:70–95

Carlton DW (2010) Revising the horizontal merger guidelines. J Compet Law Econ 6(3):619–652

Coate MB, Fischer JH (2008) A practical guide to the hypothetical monopolist test for market definition. J Compet Law Econ 4(4):1031–1063

Department of Justice and Federal Trade Commission (2010). Horizontal merger guidelines. Available at https://www.ftc.gov/sites/default/files/attachments/merger-review/100819hmg.pdf

Ekelund R, Ford G, Jackson J (1999) Is radio advertising a distinct local market? An empirical analysis. Rev Ind Organ 14(3):239–256

Elizalde J (2013) Market definition with differentiated products. A spatial competition application. Eur J Law Econ 36(3):471–521

Elzinga K, Hogarty T (1972) The demand for beer. Rev Econ Stat 54(2):195–198

Elzinga K, Hogarty T (1973) The problem of geographic market delineation in Antimerger suits. Antitrust Bull 18:45–81

Farrell, J. and C. Shapiro (2010): Antitrust evaluation of horizontal mergers: an economic alternative to market definition. J Theor Econ, 10(1), Article 9

Forni M (2004) Using stationarity tests in antitrust market definition. Am Law Econ Rev 6(2):441–463

Harris BC, Simons JJ (1989) Focusing market definition: how much substitution is enough. Res Law Econ 12: 207–226

Horowitz I (1981) Market definition in antitrust analysis: A regression-based approach. South Econ J 48(1):1–16

Hosken D, Taylor CT (2004) Discussion of 'using stationarity tests in antitrust market definition'. Am Law Econ Rev 6(2):465–475

Ivaldi M, Lörincz S (2011) Implementing relevant market tests in antitrust policy: application to computer servers. Rev Law Econ 7(1):29–71

Johnson FI (1989) Market definition under the merger guidelines: critical elasticities. Res Law Econ 12: 235–246

Kamerschen DR (1994) Testing for antitrust market definition under the Federal Government guidelines. J Leg Econ 4(1):1–10

Kaplow L (2010) Why (Ever) define markets? Harv Law Rev 124:437–517

Kaplow L (2011) Market definition and the merger guidelines. Rev Ind Organ 39:107–125

Marshall A (1920) Principles of economics, book 5. MacMillan, London

O'Brien DP, Wickelgren AL (2003) A critical analysis of critical loss analysis. Antitrust Law J 71(1):161–184

Scheffman DT, Spiller PT (1987) Geographic market definition under the U.S. Department of Justice merger guidelines. J Law Econ 30:123–147

Slade ME (1986) Exogeneity tests of market boundaries applied to petroleum products. J Ind Econ 34(3): 291–302

Stigler GJ, Sherwin RA (1985) The geographic extent of the market. J Law Econ 28(3):555–586

Uri ND, Rifkin EJ (1985) Geographic markets, causality and railroad deregulation. Rev Econ Stat 67(3): 422–428

Uri ND, Howell J, Rifkin EJ (1985) On defining geographic markets. Appl Econ 17(6):959–977

Werden GJ (1990) The limited relevance of patient migration data in market delineation for hospital merger cases. J Health Econ 8(4):363–376

Werden GJ (1998) Demand elasticities in antitrust analysis. Antitrust Law J 66(2):363–414

Werden GJ (2003) The 1982 merger guidelines and the ascent of the hypothetical monopolist test. Antitrust Law J 71(1):253–275

Werden GJ, Froeb LM (1993) Correlation, causality and all that jazz: the inherent shortcomings of price tests for antitrust market delineation. Rev Ind Organ 8(3): 329–353

# Market Failure: Analysis

José Luis Gómez-Barroso
Dpto. Economía Aplicada e Historia Económica,
UNED (Universidad Nacional de Educación a
Distancia), Madrid, Spain

## Abstract

Given that there is no agreement on the procedure by which economic efficiency should be measured, a closed catalogue of market failures cannot be talked about. There is however a reasonable agreement in economic literature on the identification of up to a total of five reasons for the existence of market failures: public goods, externalities, imperfect competition, information failures, and incomplete markets. There are three other situations that some authors also include in the list of market failures: merit goods, an unbalanced macroeconomic situation, and economic situations that assault criteria of equity.

## Definition

Market failure is any situation in which the autonomous action of the market does not lead to an economically efficient outcome.

## Introduction

Drawing the border between public and private is and has been a constant concern throughout the history of human thought. The economy is one of the fields in which such a distinction is vital. In this discipline, whatever space there is for what is public is derived from the answer to a basic question that every economist has faced: does the joint action of individual agents in pursuit of self-interest (utility or "happiness" in the case of individuals; benefit in the case of entrepreneurs) cause an acceptable result from the point of view of the common interest? When the answer is no, public intervention to correct such situations could be admissible. The unacceptability of the result of private activity derives from the use of equity or efficiency criteria. In this second case, that is, when the market does not lead to an efficient outcome of its own accord, it is said that we are in the presence of market failure.

Given that there is no agreement on the procedure by which economic efficiency should be measured (not even on the actual definition of the concept of efficiency), a closed catalogue of market failures cannot be talked about. Consider that some school of economic thought even denies their existence. Nor is there consensus on what should be done (or even whether anything should be done) to correct market failures.

By dint of being an entry in an encyclopedia proceeds to adopt a nonrestrictive approach, and the following section describes all the circumstances in which the existence of a market failure has been reasonably argued. There is insistence on the fact that many currents of economic thought would reduce the list and would furthermore qualify the extension that should be given to each type of market failure. In this respect, the definitions given are the most repeated and accepted, though for each market failure there is an extensive bibliography that would enable each of the concepts to be specified, formalized, or criticized.

## Types of Market Failures

There is a reasonable agreement in economic literature on the identification of up to a total of five reasons for the existence of market failures. They are not mutually exclusive or even independent. For the purposes of classification, they can be regrouped into two blocks:

- Those linked to the characteristics of the activity or of the good in itself: public goods and externalities
- Those related to the market situation: imperfect competition, from which information failures can become independent, and incomplete markets

There are three other situations that some authors also include in the list of market failures, though this is not the norm in literature:

- Merit goods
- An unbalanced macroeconomic situation (existence of unemployment, waste of resources)
- Economic situations that assault criteria of equity

### Intrinsic Characteristics of the Good or Activity

#### Public Goods

There are two characteristics that define a public good: it is non-excludable and consumption is non-rivalrous. In other words, no one can be deprived of enjoying the good and its consumption by an individual does not exhaust it or even affect the utility that others can extract from its consumption. The most typical examples that are mentioned in literature are national defense or coastal lighthouses.

"Pure" public goods that rigorously meet these two restrictions are few and far between. A large part of those considered public goods possess these attributes under certain conditions and could even be treated as private in different contexts. Moreover, the situations in which it is possible to talk about public goods are sometimes circumstantial. In a free wireless Internet area, consumption is non-rivalrous, as long as there are no agglomerations, but if the number of connections is excessive, "consumption" of the good by other people reduces utility and there is competition for the resources. The term impure public goods is usually used when there is a congestion problem.

With some goods, even though their underlying structure functions *naturally* as a public good, uses that break the public space in private spheres, where access is conditioned, can be developed. If a beach becomes private, a toll is set up on a motorway, or a television transmission is codified, it is evident that exclusion is possible. These situations are described as club goods.

In the case of public goods, the non-efficient outcome originates in the market not being interested in offering these goods, since in a situation of impossible exclusion income would depend on the will to contribute to their financing by those who enjoy them: how can the "right to see" a fireworks display be charged? And if exclusion can be and is actually chosen, non-efficiencies would also be generated, considering that the marginal cost of the enjoyment experienced by an additional person is zero (non-rivalrous consumption). The second reason why a non-efficient situation can be created is the overexploitation of goods that would otherwise be public. Excessive use does not only end non-rivalry in consumption, but can lead, in the extreme, to the disappearance or exhaustion of resources that are common property, such as fishing grounds or aquifers.

Nowadays, the concept of public goods has been extended from a local scale to a global scale with the introduction of the concept of global public goods. World peace, financial stability, and the eradication of epidemics would be global public goods because, once achieved, their benefits, which no one could be left out of, would be geographically unlimited. What is true is that some of these concepts, such as peace, are more desirable political objectives than goods that could be supplied by the market and, therefore, it is not strictly accurate to talk about market failures.

### Externalities

There is an externality when a concrete activity influences other activities or individuals that do not directly participate in the first activity. Externalities can be positive, if this spillover is beneficial, or negative, if they cause harm. The most typical examples collected in literature are, respectively, fruit trees pollinated by bees from a nearby apiarist and the contamination of a river.

In the presence of externalities, a social cost should be added to the internal cost reflected in a company's bookkeeping. The opening of several drinking bars in a specific neighborhood may favor certain businesses in the area (car parks, takeaway restaurants), but may have a negative effect on local residents being able to sleep. Not considering this social cost (both positive and negative) will lead to the production of an amount of the good that is greater or less than what is socially desired, resulting in a non-efficient allocation of resources. This could be resolved if the parties involved negotiated (and, therefore, the externalities "become internalized"), but even if it were possible to locate all the potential beneficiaries or injured parties, it would be difficult to reach agreements, and, moreover, were agreements to be reached, the transaction costs could exceed the benefits generated by the elimination of undesired external effects.

Nowadays, so-called network externalities associated with the growth in the number of users who use a service are gaining importance. There are direct and indirect network externalities. The first arise from the fact that each new subscriber benefits from access to the group of preexisting users, but at the same time assumes a new possibility for communication (real or potential) for this customer base already connected. The second arise from an increase in the quality or quantity of available services, a catalogue that grows with the number of users.

## Market Situation

### Failure of Competition

The prevailing opinion in economics is that perfect competition is the market structure that leads to efficient outcomes. The problem is that markets with perfect competition are a fiction, since the conditions required to receive this description are impossible to achieve in practice. There are several types of obstacles to perfect competition: differentiated (not uniform) products, lack of information, producers of an influential size that use their power to hinder the actions of their rivals, need for prior investment or other entry barriers. Therefore, in practice almost all markets have imperfect competition. Or the other way round, it could be interpreted that a market failure would be the outcome of almost all markets.

As the catalogue of possibilities is extremely extensive and covers any situation between perfect competition and a monopoly, the problem lies in deciding when the failure of competition is considered sufficiently significant to assume that it generates non-efficiencies. Furthermore, markets are dynamic and situations change. It is in this respect important to determine what constitutes a "reasonable" waiting period before assessing whether or not obstacles to developing "sufficient" competition are disappearing.

Natural monopolies are included in this category of competition failures, even though on some occasions they are mentioned as independent market failures. It is important to stress that it is not market structure per se but the result to which this structure can lead that generates the market failure: if only one bicycle shop existed in a town with a population of 5,000 (in which "there is no space" for two shops), this would not be a problem; the problem would be that its owner would take advantage of his condition as the only supplier to set abusive prices or conditions.

### Incomplete Markets

In this case, the problem is not that market structure is far from perfect competition. It is simply that no one provides the service to certain users. Properly speaking (in terms of efficiency), it is only possible to talk of incomplete markets when there is a demand not met by producers, and the cost of satisfying this demand is lower than what those seeking the products would be willing to pay.

### Information Failures

Strictly speaking, information failures are another of the causes that contribute to imperfect competition. The point of considering that it is an autonomous market failure comes from its special importance and from the fact that it also affects the demand side, unlike other "imperfections" linked to the offer.

If the information is incomplete or scant, the producer could not use the most suitable factors or reach all potential customers; in turn, the consumer could not choose the product or supplier that most suits him. One particular case is that of the "experience goods," in which it is necessary to have had a prior experience before fully appreciating their value: the consumer's lack of knowledge can reduce potential demand.

When information is asymmetric, the parties have different knowledge of a specific fact and the best informed party may use this advantage for his benefit. Asymmetries can lead to different situations of non-efficiency. Some of them have been formalized as specific categories. Adverse selection occurs when the majority of a market is made up by goods/customers/producers that the other party would not choose having all the information (insurance is taken out by those that are more likely to need it, something the company does not know; in a used items market, most present hidden defects). A moral hazard problem may exist when someone bears the potential negative consequences of a certain action performed by someone else, an action that is not observable for the first (and therefore, for instance, individuals can assume greater risks in their decisions or even be tempted to cheat).

## Other Possible Public Goods

### Merit Goods

Merit goods are those whose consumption the State judges to be "positive" and, therefore, considers it appropriate to encourage it. Examples are education or using a seat belt in cars. Their opposites are demerit goods, such as the consumption of drugs. In merit (or demerit) goods, therefore, public opinion differs from the private assessment. There is an interest attributable to the community as a whole that does not result from the "mere" addition of individual interests. Whoever judge efficiency by assessing common or social interest in this way do then consider that we are facing a market failure.

It is not often, however, that merit goods appear in this category. For many authors, the foundation of the argument presented in the above paragraph has nothing to do with the ability or inability of the market to supply these goods efficiently. Others, even recognizing that we are facing cases of "consumer myopia" (in which consumers would not be able to assess self-interest; merit goods would become exceptions to the premise that it is the consumers themselves who are better placed to maximize their welfare according to their current income), place them in a different category of possible justification of public intervention. Finally, others consider that we are facing a case of mere presence of externalities.

### Macroeconomic Situation

The fact that macroeconomic indicators are not performing well seems to indicate that the market (understood as an abstract entity formed by the union of all specific markets of goods and services) is not operating correctly, that is, it is not achieving an efficient outcome. Specifically, in the presence of a high unemployment rate, it would be logical to think that the economy could achieve a better (more efficient) outcome if part of the now wasted resources was used.

In order that this failure *of the* market be caused, failures *in some* markets or also in the structures framing the development of economic activity (and which, therefore, affect all markets) should be produced. This feature is the reason why there is a view (not widely shared) that advocates for the existence of *a* market failure.

### Equity

Income distribution criteria are usually considered an additional cause among the reasons that could justify State activity. Some authors, however, believe that in situations in which wealth distribution is very unequal, it is also necessary to talk about market failure. Their argument is that people with scant purchasing power can barely "communicate" with the market to make it aware of their needs, since only those who can pay the prices of the goods and services offered by companies manage this. Other authors consider that achieving an equitable society (or at least the reduction of poverty) would enter into an extensive definition of public good.

## Cross-References

- ▶ Anticommons, Tragedy of the
- ▶ Economic Efficiency
- ▶ Efficient Market
- ▶ Externalities
- ▶ Government Failure
- ▶ Market Failure: History
- ▶ Public Choice: The Virginia School
- ▶ Public Goods

# Market Failure: History

José Luis Gómez-Barroso
Dpto. Economía Aplicada e Historia Económica, UNED (Universidad Nacional de Educación a Distancia), Madrid, Spain

**Abstract**

The existence of market failures is linked to any opinion on the role reserved for the State in the economy. In practice, this means that the notion of market failure (though not necessarily the term) can be traced back throughout all contributions made to economic science. This entry reviews the historical evolution of the opinion on market efficiency and, consequently, of the opinion on the existence of market failures.

## Definition

Market failure is any situation in which the autonomous action of the market does not lead to an economically efficient outcome.

## Introduction

The goal of this entry is to obtain some knowledge, albeit meager, of the different conceptual positions with regard to market failures. To that end, it reviews the historical evolution of the opinion on market efficiency and, consequently, of the opinion on the existence of market failures. It is obvious to say that, as in many branches of knowledge and specifically in economics, those that are considered landmarks are based to a large extent on previous developments and the mere description of the most significant contributions necessarily leaves other valuable works to one side.

## Historical Evolution of Economic Thought on Market Failures

Economic literature usually awards Bator (1958) the merit of having coined (or of at least having used it for the first time in a publication) the term "market failure," defined as *the failure of a more or less idealized system of price-market institutions to sustain "desirable" activities or to estop* [sic] *"undesirable" activities*.

However, the existence of market failures appears to be explicitly or implicitly linked to any opinion on the role reserved for the State in the economy. This means, in practice, that one way or another, the notion of market failure (though not necessarily the term) can be traced back throughout all contributions made to economic science. And although the work of Adam Smith, and his "invisible hand," would seem to mark, in the opinion of many, a starting point in the construction of a system capable of suitably assessing the question of efficiency in markets, what is true is that reasonings in one or another sense, with varying degrees of rigor, can be found in any chapter of the history of economic thought.

It can be generically stated that all schools prior to Smith had doubts about whether private activity (the markets) managed to reconcile individual and social welfare. The ancient Greek thinkers, who did not conceive of the economy as an autonomous discipline, saw in its fellow citizens' pursuit of wealth a danger for the harmony of social order. Consequently, both Aristotle and Plato invoked strict control of economic activity by the State. In *Politics*, Aristotle proposed a superintendency whose main functions were to see that everyone involved in transactions was honest and holds to their agreements and contracts and that orderliness was maintained; some authors have even suggested that Aristotle's idea was that these supervisors could set prices (Mayhew 1993).

With all the considerations derived from some very different context and motivation, these ideas were taken up again in the Middle Ages by scholastics to reach some similar conclusions: the sinning nature of man means that, against the will of God, his love for himself comes before his love for others and, therefore, the result of unregulated individual activity (ergo the market) does not agree with divine dictates. The intervention of the authority to ensure the harmony of socioeconomic order is, once more, the corollary of this reasoning.

For mercantilists, in the sixteenth and seventeenth centuries, it was not the precepts of justice, be it divine or not, but national interest that was discredited when self-interest was pursued. And given that the first representation of this national interest is the accumulation of precious metals, commercial activity is the sector where control of private initiative should be directed at in the first place.

As a reaction to mercantilism, in the eighteenth century, the physiocrats postulated the abolition of obstacles to trade. Even though their thinking often appears to be associated with the phrase *laissez-faire, laissez-passer*, the pursuit of self-interest, derived from an excessive demand for manufactured and luxury goods, continued to be part of the problem that had to be resolved (Medema 2009). In fact, François Quesnay, one of the greatest representatives of this school, advocated control of the markets, especially of agrarian markets, given their special transcendence due to agriculture being the source of the *produit net*.

What distinguishes Adam Smith from his predecessors is the assessment of the result that the sum of individual actions produces: even if individuals do not consider in their actions anything similar to the common good, but rather the strictly personal, the most beneficial outcome for society is derived from the sum of all these actions. It is important to stress the idea of "sum," since Smith in *Wealth of Nations* gives some 60 examples in which the pursuit of self-interest causes harmful consequences for social good (Kennedy 2009). Moreover, the idea of an invisible hand infallibly guiding the markets is more a modern interpretation of his thinking than the foundation of his work, in which he only uses the expression incidentally and in the religious and cultural context of his time (among many others Harrison 2011; Kennedy 2009). Whatever interpretation is given to the metaphor of the invisible hand, what Smith is clear about is that public interference could not improve the result of private activity. Again, here it is possible to make a (important) qualification, since the three functions assigned to the sovereign in the system of "natural freedom" described in Book IV of *Wealth of Nations* are defense, justice, and "the duty of erecting and maintaining certain public works and certain public institutions, which it can never be for the interest of any individual, or small number of individuals, to erect and maintain," an exception that has even led to calling him "cautious interventionist" (Reisman 1998).

In spite of all the detailed statements that can be made on what Smith wrote, it is at least unambiguous that the market should be much less guided by governments or religions in his system of natural freedom than in all contributions that had been made up to then. His vision was shared, and refined, by the classical economists of the nineteenth century. Without being too ardent in their defense of *laissez-faire*, as they are often represented, for them the pursuit of self-interest, duly channeled through the activity of the markets, produces results that, on most occasions, are better than any other that could be obtained by government policy. The assessment of those results and, therefore, the conclusion they reach incorporate ideas developed by Jeremy Bentham and his utilitarianism ethic.

By the middle of the century, some nuances on firmly held concepts began to be introduced. The first to do so was John Stuart Mill, who in his "harm principle" puts as a limit to individual activity the cases in which said activity negatively affected the interests of others, which in modern terminology would be called the presence of negative externalities. As with so many theorists, his thinking is more complex than what is often presented: the circumstances in which public intervention could be admissible include situations in which individuals are not able to judge the result of their own actions (e.g., with regard to education). The above notwithstanding, Mill shares his reservations regarding public intervention with classical economists (there is a clear rule for not interfering, but none for interfering), though these reservations come more from his misgivings regarding the competence of governors than from some solid theoretical principles. Henry Sidgwick extended even further the catalogue of situations in which the principle of *laissez-faire* did not maximize common welfare. Examples are the overexploitation of natural resources, the occasions on which companies do not offer sufficient quantities of goods or services because they cannot recoup their investment or the cases in which there is not enough information on the effects of a certain product or action. For Sidgwick, a direct need for public intervention did not, however, come about in these situations. His answer to the question follows the rules of utilitarianism to their extreme and is, therefore, much more pragmatic than that of Mill: the cost of potential intervention (that includes aspects that already concerned many of his predecessors, such as corruption and the possibility that certain groups are intentionally favored) should also be valued and then confronted with the potential benefit obtained.

At the dawn of a new century, economists from the Cambridge School applied the tools of marginalism to the problem of market limits. Alfred Marshall, by means of the consumer surplus calculation, identified situations in which public activity could increase welfare. Arthur Cecil Pigou, comparing social and private net marginal product, gave much more concrete form to what we today call the theory of externalities. His book *The Economics of Welfare* (Pigou 1920) became an obligatory support or criticism in any subsequent contribution to the debate. The role reserved for the State was, according to Caldari and Masini (2011) and despite what is usually argued, more important for Marshall (for whom the market should be substituted in questions of social relevance) than for Pigou (for whom it should merely be complemented). All this in theoretical terms, since in practice (normative economics against positive economics) both authors, especially Marshall, did not openly commit to intervention given the limitations and inefficiencies that both pointed out in the political processes (Backhouse and Medema 2012).

In the following decades, these misgivings disappeared to the extent that economists who extended the work of Pigou not only kept enhancing and shaping the theory of externalities but made progress in the mathematical demonstration of the benefit generated by public intervention (therefore necessary) in these situations (Meade 1952; Scitovsky 1954; Buchanan and Stubblebine 1962). Also, by the middle of the century, further progress in the categorization of public goods was made; the work of Samuelson (1954), where the characteristics of "collective consumption" goods are described, and that of Buchanan (1965), on impure public goods, merit to be highlighted.

But at the same time as this current was developing, other economists were challenging their conclusions: the critical work of Coase marks an inflection point accompanied in time by the theory of Government failures.

Though sketched in other previous works, it was in *The Problem of Social Cost* (Coase 1960) that Ronald Coase structured his arguments. For him, when an activity is restricted due to the presence of negative externalities, there is also a cost associated with the restriction of such activity that is not taken into account in the calculations of social and private welfare. Which of the two “damages” is permitted is a question of assigning property rights. That assignation, though, does not necessarily lead to an efficient outcome. Ideally, it is not public intervention, but rather negotiation (the market) that could lead to an optimal (efficient) situation if the rights were well defined and there were no transaction costs. When this is not fulfilled, other options are possible. Regulation is one of them. However, if the costs associated with regulation exceed those it aims to prevent, it would leave not doing anything as the only solution; thus, the importance of having an appropriate institutional structure.

It was also in the second half of the century that a theory that did not strictly refer to market failures, but which should compulsorily be mentioned, was formulated: it is the parallel theory of Government failures. The precedents of the school of public or collective choice, as they are known, can be traced in the work of the Italian school of *scienza delle finanze* and of Knut Wicksell, who had incorporated public decision processes into theoretical analyses, thereby turning them into a factor that also determines what should be the nature and extension of the functions of the State (see Medema 2009), and further on in the work of Kenneth Arrow (most especially in Arrow 1951). With this foundation, some economists from the universities of Virginia and Chicago in the early 1960s developed a theory whose maxim was that individuals who choose between alternatives are guided by an actual rationality, both when they do so in the public function as when they choose it for themselves: they always tend to maximize self-interest.

Further on in the second half of the twentieth century, some developments came about in the theory of market failures, very linked overall to the exploration of the consequences of the asymmetries of information, which includes concepts such as adverse selection (the famous market for “lemons” described in Akerlof 1970), the moral hazard, or the lock-in effects. There has also been abundant literature theoretically or empirically resisting these advances: a dozen studies are compiled in Cowen and Crampton (2002).

At this point, it is apparent that some important schools have been left out of this historical summary. Though they have not directly referred to market failures or even to self-interest, their concept of the role of the State enables us to infer what their position is in this regard. Here we consider three that are basic for understanding modern economic thought: Marxism, the Austrian School, and Keynesianism.

For Marxism, the market always generates undesired results and, therefore, does not consider perfect markets (in which there is also capitalist exploitation and economic crises) as a desirable or reasonable end. Therefore, market failures are an irrelevant argument or, from another perspective, all markets are pure failure.

The spontaneous order advocated by the Austrian School generates a more efficient allocation of society’s resources than that which any design can achieve. There are, therefore, no market failures, or rather it would be impossible to know whether the market is failing. To do so it would be necessary to carry out an impossible assessment, since they deny the neoclassical concept of efficiency that is substituted by the non-hindrance of the actions of individuals. In this respect, any market failures would come from public action.

Finally, Keynesianism places the emphasis on the stickiness of prices and (especially) of wages in the short run, a circumstance that makes markets with no intervention generally not able to generate efficient outcomes.

## Cross-References

- ▶ Coase and Property Rights
- ▶ Market Failure: Analysis

- ▶ Mercantilism
- ▶ Smith, Adam

## References

Akerlof GA (1970) The market for lemons: quality uncertainty and the market mechanism. Q J Econ 84(3):488–500

Arrow KJ (1951) Social choice and individual values. Wiley, New York

Backhouse RE, Medema SG (2012) Economists and the analysis of government failure: fallacies in the Chicago and Virginia interpretations of Cambridge welfare economics. Camb J Econ 36(4):981–994

Bator FM (1958) The anatomy of market failure. Q J Econ 72(3):351–379

Buchanan JM (1965) An economic theory of clubs. Economica 32(125):1–14

Buchanan JM, Stubblebine WC (1962) Externality. Economica 29(116):371–384

Caldari K, Masini F (2011) Pigouvian versus Marshallian tax: market failure, public intervention and the problem of externalities. Eur J Hist Econ Thought 18(5): 715–732

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

Cowen T, Crampton E (eds) (2002) Market failure or success – the new debate. Edward Elgar, Cheltenham/ Norhampton

Harrison P (2011) Adam Smith and the history of the invisible hand. J Hist Ideas 72(1):29–49

Kennedy G (2009) Adam Smith and the invisible hand: from metaphor to myth. Econ J Watch 6(2): 239–263

Mayhew R (1993) Aristotle on property. Rev Metaphys 46(4):803–831

Meade JE (1952) External economies and diseconomies in a competitive situation. Econ J 62(245):54–67

Medema SG (2009) The hesitant hand. Taming self-interest in the history of economic ideas. Princeton University Press, Princeton/Oxford

Pigou AC (1920) The economics of welfare. Macmillan and Co., London

Reisman DA (1998) Adam Smith on market and state. J Inst Theor Econ 154(2):357–383

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

Scitovsky T (1954) Two concepts of external economies. J Polit Econ 62(2):143–151

## Marking

- ▶ Labeling

## Mass Media

- ▶ Media

## Mass Tort Litigation

- ▶ Class Action and Aggregate Litigations

## Mass Tort Litigation: Asbestos

Michelle J. White
University of California, San Diego, La Jolla, California, USA

### Abstract

Litigation over harm due to asbestos exposure is the largest mass tort in U.S. history. This article explores why the asbestos mass tort grew so large and argues that a large set of factors, rather than a single cause, are needed to explain the size of the asbestos mass tort. Among these factors are the serious of asbestos diseases and the fact that harm due to asbestos exposure was concealed by producers of asbestos products, changes in the law and legal procedures that favored plaintiffs, the large number of both potential plaintiffs and potential defendants, and plaintiffs' lawyers high profit from finding and representing asbestos claimants. I also argue that because of the unique circumstances explaining asbestos litigation, future mass torts are unlikely to be as large. The article also discusses research on asbestos litigation and explores various approaches–successful and unsuccessful–that have been proposed to reduce the number of asbestos claims.

### Definition

Mass Tort: A mass tort involves numerous plaintiffs filing civil lawsuits against one or a few corporate defendants in state or federal court. The plaintiffs allege that they were harmed by exposure to products produced by the defendants. Lawsuits may or may not be grouped in a class action. Law firms representing plaintiffs in mass torts often use advertising to locate and recruit plaintiffs.

Asbestos litigation is the largest mass tort in US history. As of 2002, 730,000 people had filed lawsuits against more than 8,400 defendants, and the cost of resolving claims was estimated at $70 billion. The number of claims increased fourfold in the 1990s, and, in 2000 alone, 12 large companies reported that 520,000 new claims were filed against them. Because individual plaintiffs typically sue many defendants, estimates of the total number of asbestos claims range as high as 10 million. As of 2003, 73 corporations had gone bankrupt due to asbestos liabilities (Carroll et al. 2005). Asbestos litigation has been extremely profitable for lawyers, since 57% of spending goes to lawyers' fees (Carroll et al. 2003). Two studies in 2001 predicted that asbestos litigation in the USA would eventually cost $200 billion (Angelina and Biggs 2001; Bhagavatula et al. 2001).

Asbestos was once considered to be a "miracle mineral" for its effectiveness as insulation and in preventing the spread of fires. It was used in ships, buildings, and consumer products, including wallboard, roofing, flooring, pipes, automotive brakes, hair dryers, children's toys, clothing, paper, and gardening products. Asbestos was used to coat the steel girders of skyscrapers such as the World Trade Center in New York, to insulate furnaces, and to make theater curtains fire resistant so that backstage fires would not spread to the seating area. Because asbestos had so many uses, estimates of the number of people who were exposed to it range from 27 to 100 million (Biggs et al. 2001).

But asbestos crumbles into microscopic fibers that become airborne and embed themselves in the lungs, causing a variety of diseases. Mesothelioma is cancer of the pleural lining around the chest and abdomen and is quickly fatal. Asbestosis is scarring of the lungs that reduces breathing capacity; it can range from non-disabling to fatal. These two are "signature diseases" that are uniquely associated with asbestos exposure. Other asbestos diseases include lung cancer, gastrointestinal cancer, and pleural plaque, which is non-disabling thickening of the pleural lining. These latter conditions can be caused either by asbestos exposure or by other factors, such as smoking. Most asbestos diseases have a long latency period, so that they do not develop until 20–40 years after exposure. Individuals' likelihood of developing asbestos disease is low, but increases as the length and intensity of exposure rise (Carroll et al. 2003).

Asbestos exposure was recognized to be harmful as early as the 1920s and safe substitutes for many of its uses were developed in the 1930s. But it nonetheless became widely used – US consumption of asbestos grew from 100,000 metric tons in 1932 to 750,000 in 1994 (Castleman 1996, p. 788). Since then, asbestos use has fallen nearly to zero, but new cases of asbestos disease continue to occur because of the long latency period.

One question concerning asbestos is why government regulation did not prevent it from becoming so widely used. The British government began in the early 1930s to regulate workplace safety in the asbestos industry and provide workers' compensation to those disabled by asbestos exposure. In the USA, many states set up workers' compensation programs around the same time. However workers' compensation programs were oriented toward providing compensation for immediate workplace injuries, while asbestos exposure caused diseases that developed many years later and were not initially connected with asbestos exposure. Because statutes of limitation were short, most workers no longer qualified for compensation at the time they developed asbestos disease.

Workers' compensation systems also protected asbestos producers from liability for harm to their workers, since these systems were workers' exclusive remedy against their employers for workplace-related harm. Thus injured asbestos workers did not qualify for workers' compensation and also were barred from suing their employers for damage. And because employers were not liable for asbestos-related harm to their workers, they had little incentive to improve workplace safety.

Workplace and product safety regulation also failed to protect workers who were exposed to asbestos. Occupational safety programs started in many US states in the 1950s and 1960s, but rules were often voluntary and poorly enforced. Some regulations actually increased workers' exposure to asbestos, such as building code regulations that required ventilation systems to be lined with asbestos insulation. As the insulation aged, it crumbled into microscopic fibers and fans blew the fibers through the workplace, where workers breathed them. Federal regulatory agencies such as the Occupational Health and Safety Administration (OSHA) and the Consumer Product Safety Commission (CPSC) came along in the 1970s and began to regulate asbestos exposure. But for many years, OSHA's workplace standards for preventing asbestos exposure were not tight enough to prevent workers from developing asbestos disease. Similarly, the CPSC's standards for limiting asbestos in consumer products in the 1970s and 1980s were mainly voluntary. Overall, state and Federal efforts to limit exposure to asbestos in the USA largely failed until the 1990s. This failure of regulation meant that many asbestos workers and product users suffered injuries due to asbestos exposure. This failure of regulation was not unique: other countries were no more successful in preventing asbestos exposure and they were generally slower than the USA to remove asbestos products from the market (White 2004; Wikipedia 2014).

In the next sections, I consider various factors that explain why asbestos litigation in the USA grew so large. I also review research on asbestos litigation and discuss various solutions – successful and unsuccessful – that have been proposed to resolve asbestos litigation.

## Why Asbestos Litigation Grew

A combination of factors, rather than a single factor, was responsible for the growth of asbestos litigation. Because workers' compensation systems are workers' exclusive legal remedy against their employers for on-the-job injuries, asbestos producers in the USA were not liable when their workers developed asbestos-related diseases. But asbestos producers were not shielded from liability to users of their products, and asbestos litigation therefore developed based on product liability law. The first successful trial of a lawsuit for damage due to asbestos exposure occurred in 1973 and involved an insulation worker who sued one of the large manufacturers of asbestos insulation (Borel v. Fibreboard, 443 F.2nd 1076 [5th Cir. 1973]). During the ensuing decade, 25,000 additional lawsuits were filed against asbestos product manufacturers and the number of lawsuits continued to grow. Because asbestos lawsuits were brought under products liability law rather than workers' compensation, plaintiffs could receive both compensatory and punitive damage awards. Damage awards could be in the millions of dollars, especially when juries awarded punitive damages (Berenson 2003).

One factor that favored asbestos plaintiffs was a change in products liability law in the 1960s that made producers strictly liable for harm to users of their products; previously, they were liable only if they were found to be negligent. The strict liability doctrine made producers liable as long as their products were "unreasonably dangerous," or users were not adequately warned of the danger. The change from negligence to strict liability made it easier for plaintiffs to win asbestos lawsuits, both because asbestos products were extremely dangerous and because they rarely contained warnings.

Another factor that favored plaintiffs in asbestos litigation is that a number of plaintiffs' law firms specialized in handling asbestos claims. These law firms invested in developing evidence against asbestos producers that could be used in all of their lawsuits. The need for law firms to invest in developing evidence kept the number of entrants small, so that the asbestos litigation "industry" remained concentrated, with the ten top law firms representing 50–75% of asbestos claims filed. The high concentration meant that profits were high (Carroll et al. 2003).

In developing a strong legal case against asbestos manufacturers, plaintiffs' lawyers were aided by the fact that several independent epidemiological studies were published in the 1960s that demonstrated strong links between asbestos exposure and asbestos disease. Asbestos plaintiffs' lawyers also developed evidence that producers conducted research on the health effects of asbestos exposure starting in the 1930s and found that exposure was harmful. But producers kept the results secret and did not warn workers or product users of the danger. This evidence of a cover-up of the dangers of asbestos exposure strengthened plaintiffs' claims in subsequent asbestos trials (Carroll et al. 2003).

The evidence suggesting a cover-up of the dangers of asbestos caused juries to frequently award punitive damages as well as compensatory damages in trials involving asbestos claims. One-sixth of all damage awards in asbestos lawsuits include punitive damages – a high proportion compared to other types of litigation. Unlike compensatory damages, punitive damage awards are often not covered by defendants' products liability insurance. The high damage awards and lack of insurance coverage made defendants eager to settle rather than litigate asbestos claims. But when claims frequently settle, they are very profitable for plaintiffs' lawyers to file, since lawyers' costs occur mainly at trial. This meant that plaintiffs' lawyers had an incentive to locate and file as many claims as possible.

Asbestos plaintiffs also benefit from the fact that they can sue many defendants. Typical asbestos plaintiffs sue 25 or more defendants, including producers of all of the asbestos products that they might have been exposed to while working or engaging in other activities. In a number of states, joint and several liability applies, so that each defendant found liable for damages is liable for the full amount of the damage award. Joint and several liability makes damage awards more valuable, since plaintiffs can collect up to the full amount of the award from any defendant(s) or their insurers. Thus even if some defendants pay little or nothing, damage awards can be collected from other defendants.

Another advantage that plaintiffs have in asbestos litigation is that their lawyers choose the most favorable court in which to file lawsuits – a phenomenon known as "forum-shopping." Plaintiffs' lawyers handling asbestos claims have a choice between filing in Federal versus state courts, and, if the latter, they can choose a state that has pro-plaintiff laws and legal procedures. Particular states are often favored because they do not require judges to approve lawyers' fees when claims are settled (this means legal fees can be higher), because they use joint and several liability and/or because they do not limit the size of punitive damage awards.

Within a particular state, plaintiffs' lawyers also choose a favorable location in which to file claims. Many asbestos claims are filed in out-of-the-way county courts where plaintiffs' lawyers have a relationship with local judges. These judges can help plaintiffs' lawyers by reducing defendants' ability to conduct pretrial discovery, scheduling trials at short notice so that defendants' lawyers have difficulty getting to the court in time, directing juries to consider awarding punitive damages, and pressuring defendants to settle. In return, plaintiffs' lawyers contribute to judges' reelection campaigns and benefit the local region by bringing in economic activity that raises demand for local hotels and restaurants. Favored locations for asbestos litigation in the past have included Madison, Illinois, Kanawha, West Virginia, and Jefferson County, Mississippi, as well as larger cities such as Philadelphia, Houston, and San Francisco – the latter because they are home to large shipyards and many former sailors who were exposed to asbestos while serving on navy ships.

Judges also developed new legal doctrines and legal procedures that favored plaintiffs and therefore encouraged plaintiffs' lawyers to file claims. One important change was a decision that greatly increased insurers' liability to asbestos claimants by legally reclassifying products liability insurance policies as premises insurance policies. While products' liability policies have a coverage limit that limits insurers' total liability under the policy to a fixed dollar figure, premises policies apply the coverage limit to each occurrence – where each individual asbestos claim is interpreted as an occurrence. Other legal changes expanded insurers' liability for claims made after the time period when their policies were in effect. These changes greatly increased insurers' liability for asbestos damage by reviving old insurance policies that had already paid out their coverage limits (Epstein 1984; Anderson 1987).

Another legal change was that judges allowed multiple asbestos lawsuits to be litigated together, thus creating informal class actions. Asbestos plaintiffs' lawyers initially tried to have all asbestos claims certified as a class action in Federal court and settled all at once, but the Supreme Court overruled two settlements of class actions involving asbestos claims (Amchen Products v. Windsor, 117 S.Ct. 2231 (1997) and Ortiz v. Fibreboard Corp., 119 S.Ct. 2295 (1999)). After these two decisions, plaintiffs' lawyers shifted to filing most asbestos claims in state courts. Judges in these courts allowed groups of lawsuits to be consolidated for either the pretrial or the trial stages of litigation, or both, using a procedure known as mass joinder. These consolidations often combined multiple claims by out-of-state plaintiffs with a small number of claims by in-state plaintiffs. The total number of claims consolidated ranged from a few to up to 9,600. Judges would hold a single trial before a single jury for all claims, with the jury sometimes making separate decisions for each plaintiff and sometimes making a single decision for all plaintiffs (Carroll et al. 2005). Combining multiple lawsuits for litigation benefits plaintiffs by making the trial outcomes more positively correlated. This makes going to trial more risky for defendants, because losing many cases at once could exhaust their insurance coverage and force them to file for bankruptcy. The more claims that are combined, the more bargaining power plaintiffs' lawyers have. Thus when large numbers of asbestos claims are consolidated for trial, defendants are likely to settle even claims that are legally weak.

Another legal change that benefitted plaintiffs in asbestos lawsuits is the use of bifurcated or reverse bifurcated trials. In a bifurcated trial, evidence concerning liability is presented first and the jury decides separately on each defendant's liability. Then the trial is suspended while plaintiffs and defendants who have been found liable bargain over a settlement. If they fail to settle, the trial is resumed at a later date – sometimes with the same jury – for the damages portion of the trial. In a reverse bifurcated trial, the format is the same, but damages are tried in the first stage and liability in the second stage. Bifurcation saves on trial time relative to holding a unitary trial if the parties settle after the first stage. The parties are also more likely to settle at the end of the first stage than before the trial starts, since they have some of the information that the trial will generate.

Reverse bifurcation was developed specifically for asbestos trials and is particularly thought to benefit plaintiffs. This is because plaintiffs often have severe damage from their asbestos exposure – making damage awards high. In contrast, plaintiffs' claims are often weak on the liability side, because they cannot show that they were exposed to particular defendants' asbestos products. So using reverse rather than straight bifurcation strengthens plaintiffs' bargaining power in settlement negotiations, because the information generated by the first stage of the trial is very favorable to plaintiffs and raises their bargaining power in settlement negotiations.

Bouquet trials are another procedural innovation developed for asbestos litigation. In a bouquet trial, a small group of asbestos plaintiffs is selected for trial from a larger group of consolidated claims. The trial group includes plaintiffs with severe asbestos disease and plaintiffs with no impairment. The idea of the bouquet trial is that the outcomes at trial for the various types of plaintiffs will be used as a template for settling the remaining claims in the larger group. Using a bouquet trial allows larger numbers of claims to be consolidated, since a bouquet trial can be held even when the full consolidated group of claims is too large to hold a single trial. One well-known example is a trial of 12 asbestos claims in Mississippi that were selected from a larger group of 1,738 asbestos claims. At the bouquet trial, the jury awarded plaintiffs damage of $4 million each. The prospect of the jury assessing similarly high damage awards for the remaining plaintiffs caused defendants to settle all the remaining claims on very favorable terms (Parloff 2002).

Another factor that allowed the asbestos mass tort to grow so large is the large number of potential plaintiffs. As discussed above, the widespread use of asbestos meant that millions of people were exposed. Typical plaintiffs include ex-sailors who were exposed to asbestos on ships during World War II, workers who install insulation, workers in shipyards and steel mills, and textile workers who were exposed to airborne asbestos fibers in factories. Plaintiffs' lawyers search for new plaintiffs by extensive advertising and by conducting mass screenings. A frequent procedure was to bring a van equipped with an X-ray machine to a factory and take chest X-rays of all the factory workers. Any found to have scarring or thickening of the lungs or the pleural lining would be signed up as asbestos plaintiffs. Doctors often read hundreds of X-rays per day and found that nearly all of them had asbestos-related damage. More recently, plaintiffs' law firms have shifted to television advertisements to recruit plaintiffs whose exposure to asbestos may be non-work-related.

Another issue that has allowed asbestos litigation to become so large is that claims are valuable even when plaintiffs have no impairment from their asbestos exposure or had no asbestos exposure so that their claims are downright fraudulent. Because asbestos lawsuits are mainly settled rather than tried, non-impaired and fraudulent claims are valuable because they increase the size of consolidations and raise plaintiffs' lawyers bargaining power with defendants. Settlements cover both fraudulent and valid claims. Estimates suggest that as few as 10% of plaintiffs with asbestos claims have asbestos-related cancers – a widely used measure of disabling asbestos disease (Carroll et al. 2003). Legal standards that allowed non-impaired plaintiffs to collect damages are an important feature of asbestos litigation.

The asbestos mass tort also involves many types of defendants. In the first stage of the litigation, defendants were the major producers of asbestos insulation. These companies eventually went bankrupt. In the second stage, these defendants were replaced by producers of asbestos-containing products, retailers that sold these products, and firms that operated workplaces containing asbestos. Examples include the automobile companies, sued because car brakes contained asbestos; Sears Roebuck, sued because its stores sold asbestos-containing products; 3M Corporation, sued because it made dust masks that didn't protect users from asbestos exposure if they used the masks improperly; and Crown Cork and Seal, sued because it briefly owned a company that included a division which produced asbestos-containing insulation. Crown Cork quickly sold the division that produced insulation, but nonetheless it eventually paid out $700 million in asbestos settlements and damage awards. Both small and large firms were sued, since even small defendants have insurance. Each time new defendants were added to the litigation, previous plaintiffs filed new claims against them. Because there were so many plaintiffs and so many potential defendants, the asbestos mass tort continued to grow.

Finally, bankruptcy also played a role in encouraging asbestos litigation. Many of the large firms that produced asbestos insulation and asbestos-containing products went bankrupt due to their asbestos liabilities – the first was the Johns-Manville Corporation in 1982. When asbestos-producing firms go bankrupt, present and future damage claims against them are assigned to a trust which receives some or all of the reorganized firms' equity and uses the funds to pay compensation to asbestos victims. Congress adopted legislation defining these trusts in 1994 and required that they follow the general outlines of the Manville Trust that was set up following the Johns-Manville bankruptcy. Trusts first estimate the number and severity of future asbestos claims against them and then determine what level of compensation payments they can pay such that their funds will cover both present and future claims. Trusts payments vary with the severity of the claimant's asbestos disease and the length of exposure to asbestos. The trusts do not require that claimants show impairment from their asbestos exposure and they use quite loose standards for demonstrating exposure to the bankrupt firm's asbestos products. This was done in order to reduce transactions costs and increase the fraction of damage payments that went to claimants rather than lawyers. However the loose standards for receiving compensation caused the number of claims to increase, causing many of the trusts to cut their compensation payments. On average, claimants with no asbestos-related impairment receive a total of around $8,000 in compensation from all of the trusts, while claimants with moderate impairment receive around $19,000. Compensation trusts have paid out a total of around $17 billion to asbestos claimants (Scarcella et al. 2013).

The bankruptcy trusts encourage asbestos litigation in two ways. First, when corporations go bankrupt, their damage payments fall drastically. This encourages plaintiffs' lawyers to find new asbestos defendants to substitute for those that have gone bankrupt. The bankruptcies thus have contributed to bringing in many new corporations as defendants whose involvement with asbestos is increasingly remote. Second, although the trusts' compensation payments are relatively small, representing trust claimants is nonetheless profitable for plaintiffs' lawyers if they represent large numbers of claims. The trusts therefore encourage plaintiffs' lawyers to continue recruiting large numbers of non-impaired claimants, since the loose compensation rules allow these claimants to receive payments from many or all of the trusts.

Overall, a combination of factors is needed to explain why asbestos litigation grew so large.

## Research on Asbestos Litigation

In White (2006), I examined why judges adopt the procedural innovations used in asbestos trials and the effect of both forum-shopping and procedural innovations on trial outcomes. The procedural innovations, discussed above, are consolidation of multiple lawsuits for trial, bifurcation and reverse bifurcation, and bouquet trials.

Why do judges adopt these innovations for asbestos trials? Judges in favored jurisdictions for asbestos litigation have crowded dockets. Because it would be impossible to hold individual trials for all cases, judges favor procedures that encourage the parties to settle and therefore reduce trial time. Consolidating claims for trial is a method of reducing trial time, because only one jury must be selected and some of the evidence can be presented only once for all plaintiffs. Consolidation also increases the probability of settlement, because trial outcomes become more positively correlated and defendants therefore find it riskier to go to trial. Bifurcating trials reduces trial time relative to holding a unitary trial, because the information generated in the first phase of trial increases the probability of settlement when the parties bargain after the first phase. Finally, bouquet trials save trial time by allowing larger numbers of asbestos claims to be consolidated – if a trial is needed, then a bouquet trial can be held when the number of claims in the consolidation would otherwise make it too large for a single trial.

The study uses a dataset consisting of all asbestos lawsuits that were tried in court to a verdict on liability or damages or both between 1987 and 2003. Each observation consists of a trial of a single plaintiff's asbestos claim against all defendants. There were around 5,200 observations in the dataset, implying that less than 1% of asbestos plaintiffs' claims go to trials.

The data include the plaintiffs' alleged disease, the trial venue, the trial outcome, whether the claim was consolidated for trial and the number of claims in the consolidated group, whether the trial was bifurcated or reverse bifurcated, whether a bouquet trial was used, and the number of defendants that each plaintiff sued.

Half of all claims had individual trials, while the rest were consolidated with at least one other claim for trial. Approximately one-fifth of trials were bifurcated or reverse bifurcated and 4% were bouquet trials. Use of the procedural innovations was geographically concentrated: bifurcated trials were frequently used in Manhattan and Philadelphia, while bouquet trials mainly occurred in Mississippi. Sixty-four percent of plaintiffs were awarded compensatory damages and the average compensatory damage award (contingent on defendants being found liable) was $1.3 million in 2003 dollars; 20% of plaintiffs were awarded punitive damages and the average punitive damage award (contingent on defendants being found liable for both compensatory and punitive damages) was $1.8 million. Plaintiffs' expected return from going to trial was $1.1 million for the entire sample, with those having mesothelioma receiving around $3 million more.

To examine the effect of consolidating claims for trial on the correlation of the trial outcomes, I computed a correlation coefficient for all trials involving two plaintiffs and compared the result with the correlation coefficient for single-plaintiff trials when plaintiffs were randomly assigned in pairs. I also followed the same procedure for three- and five-claim consolidations. The results show that the correlation coefficient of expected total damages ranges from 0.84 to 0.92 in the actual groups, compared to only 0.01–0.04 in the randomly assigned groups. The results were similar if only liability or only damages are considered. These results suggest that consolidating claims for trial makes trial outcomes much more positively correlated and supports the hypothesis that going to trial in a consolidation is much more risky for defendants.

To examine the effect of forum-shopping and the procedural innovations on trial outcomes, I estimated probit regressions explaining whether plaintiffs were awarded compensatory damages and whether they were awarded punitive damages conditional on receiving compensatory damages. I also estimated Tobit regressions explaining the amount of compensatory and punitive damages, with damages set equal to zero when the plaintiff loses. Forum-shopping was found to be extremely favorable to plaintiffs, with plaintiffs' probability of receiving compensatory damages increasing by up to 30 percentage points in the most favorable jurisdictions relative to the most commonly used jurisdiction. Also plaintiffs' probability of being awarded punitive damages rose by up to 91 percentage points in the most favorable jurisdiction relative to the most commonly used jurisdiction.

Use of the procedural innovations also increased plaintiffs' expected return from going to trial. Having a bifurcated trial raised plaintiffs' probability of being awarded compensatory damages by 27 percentage points and raised compensatory damage awards by $924,000. Having a bifurcated trial also increased plaintiffs' expected return from going to trial by $650,000. But bifurcated trials did not significantly increase plaintiffs' probability of winning punitive damages or the size of the punitive damage award. Having a bouquet trial raised plaintiffs' probability of being awarded punitive damages and caused both compensatory and punitive damage awards to be higher. Plaintiffs' expected return from going to trial increased by $1.2 million when a bouquet trial was held. Having a small consolidated trial consisting of 2–5 plaintiffs' claims increased plaintiffs' probability of winning both compensatory and punitive damages, but was associated with lower compensatory damage awards. Surprisingly, having a larger consolidated trial of six or more plaintiffs did not significantly change plaintiffs' returns from going to trial.

Overall the results suggest that the return to plaintiffs and their lawyers from filing asbestos claims is greatly increased by forum-shopping and by plaintiffs' lawyers picking jurisdictions where judges use the procedural innovations. Although the research did not address the issue of how forum-shopping and procedural innovations affect the size of asbestos settlements, the standard economic model of settlements suggests that they mirror trial outcomes and are higher in courts where plaintiffs' expected returns from going to trial are higher (Mnookin and Kornhauser 1979). Thus forum-shopping and procedural innovations are also likely to raise the amount that defendants pay to settle asbestos claims.

## Methods of Resolving Asbestos Litigation: Hypothetical and Actual

In this section, I consider solutions for resolving asbestos litigation – including both proposed solutions that were never adopted and actual solutions that were.

One proposed solution in the 1990s was to certify a class action of all asbestos claimants. In a class action, all asbestos claims are combined in a single lawsuit and all are resolved at once, usually by a settlement. Both present and future asbestos claims are resolved. Individual plaintiffs would be bound by the outcome of the class action and would not have had the right to opt out. The Federal courts certified two class actions of asbestos claimants, but – as discussed above – the US Supreme Court rejected both class certifications in 1997 and 1999, on the grounds that asbestos claimants were too diverse to be combined into a single class.

This was followed by another proposed solution for asbestos litigation: a Federal government-administered compensation scheme for asbestos victims. The proposed bill was the Fairness Asbestos Injury Resolution or "FAIR" Act of 2005, S. 852. It was based on previous federally administered programs, one that compensated miners who developed black lung disease and one that compensated children harmed by childhood vaccines. Compensation of up to $140 billion would have been financed by levies on asbestos producers and insurers. Asbestos victims would lose their right to file lawsuits, but would instead receive compensation from the trust. Claimants who had mesothelioma or cancer would receive the highest awards of $1.1 million and those with less disabling diseases would receive $25,000 or more. Non-impaired claimants would receive medical monitoring, but no compensation (Stengel 2006). However the FAIR Act was not enacted (Barnes 2011).

While both the class action settlement and the compensation scheme for asbestos claims failed, courts began in the early 2000s to adopt new procedural innovations that reduced the amount of asbestos litigation. One such device was the "inactive docket" which put claims by non-impaired asbestos plaintiffs on an inactive basis, preserving their right to sue in the future, but preventing their claims from proceeding in the legal system until they become impaired from their asbestos exposure. Inactive dockets solve the problem that plaintiffs must file claims quickly after discovering their asbestos-related harm in order to satisfy statutes of limitations. But because most asbestos claims are classified as inactive, asbestos litigation now consists mainly of plaintiffs who have severe asbestos-related diseases.

As a result of the use of inactive dockets, plaintiffs' lawyers can no longer litigate large groups of claims consisting mainly of non-impaired plaintiffs, and they therefore have less bargaining power to force defendants to settle. The fraction of asbestos damage awards going to non-impaired plaintiffs has fallen from around 50% in 1997–1999 to less than 5% in 2013 (Scarcella et al. 2013). This change has greatly reduced plaintiffs' lawyers' return from recruiting non-impaired asbestos claimants. It has been so successful in reducing the volume of asbestos litigation that an observer is led to wonder why judges did not adopt it much earlier.

Another recent development is that some states that were centers for asbestos litigation have adopted legal reforms to discourage the filing of asbestos claims in the state. An important change in several states was to bar judges from consolidating out-of-state with in-state asbestos claims for litigation. As a result, out-of-state claims could no longer be litigated in the state and therefore plaintiffs' lawyers could no longer put together large consolidations. Among states that previously allowed large consolidations of asbestos claims, West Virginia, Mississippi, and Illinois all made changes along these lines in the early 2000s. Several other states changed their legal rules to explicitly disallow large consolidations of asbestos claims, although they generally still allow out-of-state claims to be consolidated with in-state asbestos claims. Another change is that New York, Texas, and several other states substituted proportional liability for joint and several liability to asbestos claimants, so that individual defendants are no longer liable for plaintiffs' entire damage award. This shields non-bankrupt defendants from being held liable for bankrupt defendants' share of plaintiffs' damage (Hanlon and Geise 2007). The result of these changes in state law is that most asbestos litigation now involves a much smaller number of claims by plaintiffs with serious asbestos-related diseases and these claims are litigated individually or in small groups.

Finally, judges have become more likely to dismiss fraudulent claims, rather than pressure defendants to settle them. This approach was used recently to resolve a different mass tort:

claims for damage due to silica exposure. Silica litigation was a spinoff from asbestos litigation and took a similar form. Plaintiffs allege harm from inhaling airborne silica crystals that can lead to scarring of the lung lining, silicosis, or lung cancer. Because of the similarity between asbestos disease and silica disease, plaintiffs' lawyers recruit silica claimants using the same mass screenings with chest X-rays that they use to recruit asbestos claimants. In fact, plaintiffs' lawyers often file both silica claims and asbestos claims on behalf of the same individuals, using the same chest X-rays; this is despite the fact that it is rare for individuals to have been exposed to both silica and asbestos. However the judge who presided over the silica litigation dismissed nearly all of the claims on the grounds that they were fraudulent and threatened to bring criminal charges against the doctors who read the plaintiffs' X-rays. This effectively ended the silica mass tort, leaving only a small number of lawsuits by plaintiffs with severe silica-related disease. The publicity given to the silica litigation has probably made judges more likely to dismiss asbestos claims as well (Behrens and Goldberg 2005/2006).

## Future Directions

Because asbestos litigation has been so lucrative, plaintiffs' lawyers have searched widely for other defective products that could serve as the basis for new mass torts, using the techniques they developed for asbestos litigation. Among potential future mass torts are litigation involving harm due to exposure to lead paint, harm due to guns, and claims of obesity due to consumption of fast food (White 2004). However none of these spinoff mass torts have been successful in court.

But the asbestos mass tort itself continues to mutate into new forms that keep it alive. One recent development is lawsuits filed by family members of asbestos workers who claim second-hand exposure to asbestos from relatives' clothing. Family members, unlike workers themselves, are not barred by workers' compensation from suing their relatives' employers. Thus they can both sue their relatives' employers and the producers of asbestos products that their relatives were exposed to. Another new development in asbestos litigation is claims by lung cancer victims against asbestos producers and the asbestos bankruptcy trusts. Most lung cancer is caused by smoking, but plaintiffs with lung cancer nonetheless claim that their cancer was caused by exposure to asbestos. These claims qualify for compensation from the asbestos bankruptcy trusts, and, because lung cancer is a serious disease, their lawsuits against non-bankrupt defendants are not placed on the inactive docket (Nocera 2013). And since there are 200,000 new lung cancer cases each year compared to only 2–3,000 new mesothelioma cases, lung cancer claims present a valuable opportunity for lawyers to continue the asbestos mass tort.

## References

Anderson DR (1987) Financing asbestos claims: coverage issues, Manville's bankruptcy and the claims facility. J Risk Insur 54(3):429–451

Angelina M, Biggs J (2001) Sizing up asbestos exposure. *Mealey's Litigation Report: Asbestos* 16:32–38

Barnes J (2011) Dust-up: asbestos litigation and the failure of commonsense policy reform. Georgetown University Press, Washington, DC

Behrens MA, Goldberg P (2005/2006) The asbestos litigation crisis: the tide appears to be turning. Conn Insur Law J 12:477

Berenson A (2003) 2 Large verdicts in new asbestos cases. *New York Times*, 1 Apr 2003

Bhagavatula R, Moody R, Russ J (2001) Asbestos: a moving target. AM Best Rev 102:85–90

Biggs JL et al (2001) Overview of asbestos: issues and trends. Report prepared by the American Academy of Actuaries Mass Torts Work Group. Available at www.actuary.org/pdf/casualty/mono_dec01asbestos.pdf

Carroll S, Hensler D, Abrahamse A, Gross J, White M, Scott Ashwood J, Sloss E (2003) Asbestos litigation costs and compensation. DRR-3280-ICJ. RAND Corporation, Santa Monica

Carroll S, Hensler D, Gross J, Sloss EM, Schonlau M, Abrahamse A, Scott Ashwood J (2005) Asbestos litigation. DRR-3280-ICJ. RAND Corporation, Santa Monica

Castleman BI (1996) Asbestos: medical and legal aspects, 4th edn. Aspen Law & Business, Englewood Cliffs

Epstein RA (1984) The legal and insurance dynamics of mass tort litigation. J Legal Stud XIII:475–506

Hanlon PM, Geise ER (2007) Asbestos reform – past and future. *Mealey's Litigation Report: Asbestos* 22:5, 4 Apr 2007

Mnookin R, Kornhauser L (1979) Bargaining in the shadow of the law: the case of divorce. Yale Law J 88(5):950–997

Nocera J (2013) The asbestos scam. *New York Times*, 2 Dec 2013

Parloff R (2002) The $200 billion miscarriage of justice. *Fortune*, 4 May 2002

Scarcella MC, Kelso PR, Cagnoli J (2013) Asbestos litigation, attorney advertising and bankruptcy trusts: the economic incentives behind the new recruitment of lung cancer cases. Mealey Asbest Bankrupt Rep 13:4

Stengel JL (2006) The asbestos end-game. NYU Annu Surv Am Law 62:223

White MJ (2004) Asbestos and the future of mass torts. J Econ Perspect 18:2

White MJ (2006) Asbestos litigation: procedural innovations and forum-shopping. J Legal Stud 35(2): 365–398

Wikipedia (2014) Asbestos and the law. http://en.wikipedia.org/wiki/Asbestos_and_the_law, viewed 21 Apr 2014

# Measure of Environmental Regulation

Vittoria Colombo
Università degli Studi di Torino, Turin, Italy

## Definition

Environmental law and economics has been largely focused on the effect that different environmental rules can have on economic outcomes, while only few detected the difficulty of analysis of such subject, mainly due to the measurement of environmental stringency. There is no homogeneous way to measure environmental regulation, due to its peculiarities; environmental rules are often designed based on elements other than just the legal principle, such as geographical characteristics, scientific data, the level of pollution and the industry sector involved, and elements that must be taken into account in order of not incurring in methodological mistakes. This essay examines the main approaches and indicators of environmental regulation, considering the advantages and disadvantages of any of them. It also analyzes the main risks of a wrong methodological approach to the measurement of environmental rules.

Environmental regulation is an atypical form of regulation, which needs to be assessed in its entirety and complexity. Its design requires scientific and technical information that are pivotal for reaching its objectives. Environmental rules are difficult to assess, since decisions on how to design a rule do not only depend on legal elements but also on scientific data, risk assessment, and scientific criteria.

## Environmental Stringency

Currently the most used method of measuring environmental regulation is by looking at costs imposed by it, namely, at the *environmental stringency*, which in its broad and shared definition is the "cost of polluting by firms across different sectors and policy instruments," where "a higher value represents a more stringent policy." The use of environmental stringency allows to ascertain environmental regulation as a whole, and at the same time, it is sufficiently general to allow scholars to use different methods and datasets. The main conceptual difficulties in analyzing environmental stringency can be summarized in seven broad problems: multidimensionality, simultaneity and identification, trade-off between de jure and de facto situations, industrial composition, sampling, grandfathering, and lack of data (Brunel and Levinson 2013).

## Multidimensionality

Environmental regulation is typically *multidimensional* (Wing-Hung et al. 2011): it deals with different media, such as soil, water, and air, as well as different pollutants that affect those media (dioxin, chemicals, carbon dioxide, etc.). Moreover, it has different recipients: firms, consumers, households, and even public administration itself. Finally, it can set standards or limits that vary on the basis of several factors, which can change depending on the environmental quality of territory, on the technology used by firms, or on the sector in which firms operate (Kozluk 2014a). On the one hand, there could be many environmental regulations applying to the same industry, aiming at protecting different media (water, land, human health, etc.); on the other hand, policy-makers' decisions are also relevant, with the same rules applied in different ways depending on several internal or external factors that have a link with environmental standards. The level of emissions, the location of the activity, the technology applied in a certain plant, or the sector in which the firm operates can have a decisive weight on the decision not only on the application of a certain rule but also on the intensity of the instrument itself. The intensity of the application of instruments depends on factors that are linked with the environment, and the decision to forbid a certain production method in a region could depend on specific geographical characteristics of the territory.

Multidimensionality is the reason why case studies on a specific industry sector are limited, because they are not representative of other categories not covered by that specific regulation (Sauter 2014).

## De Jure and De Facto Applications

The de jure stringency derives from environmental legislation applied in a certain legal system ex ante, while the de facto one corresponds to the real application of the de jure instruments. This assessment is made by looking at courts' decisions, namely, at the difference between regulation and its enforcement. De facto application of the rules may also involve other informal elements that are not easily detectable as judicial decisions: enforcement of rules does not only occur in courts but can take different shades, i.e., it could consist in a fine given by the administrative local authority but also in some checks made by the public authority in the concerned sites. For some environmental regulations, soft mechanisms are used, i.e., flexibility mechanisms or the provision of consultations with the authority in charge of enforcement.

## Simultaneity and Identification

Simultaneity is the difficulty of identifying the direction of causality between the rule and its effect. Countries may introduce stricter environmental regulations in order to address concerns on high-polluting industries. However, many pollution-intensive industries have consistent political power and can lobby decision-makers to enact lax environmental regulation. This makes difficult to assess the causal link between a certain environmental policy and its results (Brunel and Levinson 2013). Similar to the problem of simultaneity is the broader problem of identification, which in this case occurs when there is a problem in identifying whether a certain result is due to an environmental rule or to other regulatory instruments, i.e., legislation on employment, or to certain market characteristics (Malatu 2008).

Some scholars conducted natural experiments to avoid such problem, taking into account relevant changes in regulation forced by external factors (as a Supreme Court decision or international treaties' implementation), and comparing the *ante*- and *post*-scenarios (McConnell and Schwab 1990; Henderson 1996), others looked at instrumental variables correlated with the regulatory stringency (Xing and Kolstad 2002; Levinson and Taylor 2008) but uncorrelated with the measure of economic activity.

## Industrial Composition

Industrial composition of countries may be due to endogenous sources that are not measurable, such as the geographic characteristics or natural resources in the country, i.e., carbon or natural gas. Industrial composition, if not considered, can cause methodological problems.

## Sampling and Grandfathering

Sampling deals with the causal link assessment but related to the sample of industries subject to a given environmental policy. An industry's market share in a country can be itself the result of a determined environmental policy, i.e., high-polluting industries can be present more in a country than in another, because of that country's environmental policy toward a certain industry sector. This problem has to be taken into account when assessing the effect of policy stringency in a sector.

Also grandfathering, namely, the provision according to which new rules apply only to new situations, while the old rules continue to apply to existing situations, could create problems. In the field of environmental law, stricter rules may apply to new plants or pollution sources, while lax ones are applied to old sources of pollutions that, most probably, are also the most polluting ones. This problem concerns more consumers than firms: whereas the former ones are not bound to buy new products – consumers are not obliged to buy a new car and can continue to drive the old one, which is much more polluting than new models – the latter ones are often shifting to new technologies because they are obliged by law to do so (Heutel 2010).

## Lack of Data

Lack of data must not be intended in absolute terms, but as lack of data giving source to a robust outcome. This factor is linked with multi-dimensionality, identification, simultaneity, and sampling; often the available data consider just a small part of environmental legislation while not dealing with the assessment of the quality and quantity of all environmental regulation applied in a determined country or region.

## Approaches in Measuring Stringency

Although the definition of environmental stringency is widely accepted, it is not the same for what it concerns its *measure*.

The main challenge of measuring environmental stringency is the delineation of a conceptual framework that considers all the principal layers of environmental law and that includes them in the analysis (Sauter 2014). The distinction between market- and nonmarket-based instruments allows to assess the impacts that these two regulatory approaches may have. However, this distinction alone can be misleading and too reductive, because it does not consider other elements such as flexibility and stability of regimes (Johnstone et al. 2010). De Serres (De Serres et al. 2010) distinguishes four categories of market-based instruments: (a) environmental taxes, (b) pollution trading systems, (c) deposit-refund systems, and (d) subsidies for incentivizing environmental friendly activities. Nonmarket-based instruments include the general category of (a) command and control regulations, (b) technology-support policies, and (c) voluntary approaches.

The main approaches for measuring environmental regulation will be presented in the following.

The first approach looks at changes in a single regulation or rule. The second looks at perceptions that firm directors, public officials, and, more generally, people directly touched by environmental regulation have on its stringency. The third considers shadow prices for environmental inputs or expenditures related to environmental rules. The fourth looks at the variation of environmental performances to deduct a more or less stringent environmental legislation. Finally, composite measures aim at avoiding multi-dimensionality, by including all the elements in the same indicator.

### Change in a Single Regulation

The advantage of looking at the effect of a change in a single regulation is the level of certainty it can bring to the analysis; the risk of multi-dimensionality is highly reduced, as well as the one of simultaneity and identification. In addition, this kind of analysis is more feasible than the general one, because of the availability of data. However, the limits are relevant. By looking at the single regulation, results will be applicable only to it and not to other sectors (Smarzynska and Wei 2001). The reason is exactly multidimensionality that, when eliminated though a circumscribed analysis, limits the extension of the same results to other cases.

### Analysis Based on Perceptions

The majority of datasets on environmental stringency are based on perceptions by businessman and firms' directors or civil servants. The advantage of such approach is the availability of data and the reliability of the analysis, which can be circumscribed by asking the same questions on a limited number of regulations or media. However, the premises are not always exempted from critics. When data are used for a cross-country analysis, perceptions are difficult to compare, since the requisites for an objective perception of situations by individuals are the knowledge of the other terms of comparison, which in such case are assumed (Becker and Henderson (2000). Interviewees should be aware of the strictness of environmental regulation in other countries in order to give an objective opinion on the one of their country. Moreover, according to some authors, perception-based surveys involve some sampling problems, given by the fact that "the sample of respondents may actually be a result of environmental policies" (Becker and Henderson (2000), Botta and Kozluk (2014), p. 11).

In addition, complexity of environmental regulation makes sometimes difficult to disentangle pollution abatement costs from other costs, and sometimes for firms, it could also be beneficial to indicate wrong estimates. Finally, also the perception of the strictness of a determined regulation could be influenced by external factors and therefore be unreliable (Gallaher et al. 2006). One of such cases is when reporting is affected by business cycles: in periods of crisis, environmental regulation may be considered stricter than in other times (Brunel and Levinson 2013).

### Shadow Prices and Performance Indicator Approach

Shadow prices are the costs of polluting that can be calculated from the firm's production function.

The production function becomes the starting point for the calculation of policy stringency; emissions would be considered as an element to calculate the environmental stringency of regulation. The advantage of such approach is that, by looking directly at emissions, it considers the de facto situation and therefore assesses the real policy stringency and not the de jure one.

Performance indicators, as for the shadow price, look at the amount of emissions actually produced or at the pollution intensity in a specific media. However, contrary to the former approach, this one directly considers the amount of pollution produced as sign of environmental strictness.

Three main performance indicators have been used by the literature so far: emissions, energy consumption, and gasoline. The problem is that it cannot exclude multidimensionality and identification issues (Botta and Kozluk 2014). Most of the indicators quantify the problem to be addressed and not the stringency of environmental regulation. Identification problems are consistent, since the amount of emissions can be due to factors other than environmental regulations, which can be associated with the particular geographical structure, to the effect of other policies, and to other external factors (Albrizio et al. 2014). Also simultaneity can occur, since the amount of emissions could influence regulation itself.

### Composite Measures

Composite measures seek to introduce multidimensionality into the analysis, trying not to eliminate the inevitable complexity of environmental regulation, but including it into an index. By doing this, composite measures are more complete, because they take into account all the factors of multidimensionality (Smarzynska and Wei 2001).

However, risks are very high: given the complexity of environmental legislation, the index must be adequately constructed and fine-tuned in order avoid methodological shortcomings, which in this case can be less evident and therefore more dangerous. Another disadvantage of indexes is that they are representatives of the only de jure situation. Such shortcoming is not negligible, since the enforcement and application of environmental regulation are crucial elements of environmental stringency (Sauter 2014, p. 8).

## Choice of Indicators for Measuring Environmental Stringency

This section presents the indicators and proxies adopted by the literature for measuring stringency, namely, pollution abatement costs, the quantity of polluting emissions produced, and, finally, indexes, and then combines different measures in order to cover all possible elements of environmental stringency.

### Pollution Abatement Costs

Environmental stringency has often been defined by using the proxy of the private sector's pollution compliance costs (Keller and Levinson 2002; Levinson 1999), namely, costs that firms face to comply with environmental regulation (Carraro et al. 2010). Compliance costs can take the form of environmental taxes, liability rules, emissions limits, and technological standards. They can also include permitting costs, regulatory delays, threat of lawsuits, and the redesign of a process or product (Levinson and Taylor 2008).

Some authors considered all types of pollution abatement costs and expenditures while others only a particular media (Brunnermeier and Cohen (2003).

The principal and most used method to calculate pollution abatement costs is based on surveys collected at regional and industry level about firms' costs on abating pollution. One of the most used and complete datasets is the PACE survey (Pollution Abatement Costs and Expenditures)' which has been conducted for the first time on 1994 in the United States on annual data from 1973 and then repeated with few modifications on 1999 and 2005. It collects data on pollution abatement capital expenditures and operating costs mainly in the manufacturing industry of the United States. Other countries started collecting data on environmental expenditures. Above them are Canada, with its SEPE (*Survey of Environmental Protection Expenditures*), and the European Union with the Eurostat *Questionnaire on Environmental Protection Expenditure and Revenues* (EPE). Also the OECD has drafted a dataset on environmental protection expenditure and revenue.

### Public Entities' Pollution Abatement Expenditures

The costs of the environmental effort made by the state or other public entities to enforce environmental protection do not use directly environmental regulation, but some proxies, and were more frequent some years ago, when environmental datasets were scarce and imprecise.

Proxies for public expenditure could consist in a mix between private and public expenditures (Pearce and Palmer 2001), in the state budget for enforcement and inspections (Gray 1997), or in the number of employees in environmental agencies in relation to the number of manufacturing industries (Levinson 1999).

The advantage of such approach is that it often includes the enforcement stage, which would allow assessing also the de facto scenario. However, if the aim is to assess the impact that environmental stringency has on firms, the method of looking at public expenses, which may include also environmental actions for consumers or householders, could bring to distorted results. Second, such proxies do not have a strong causal link with environmental stringency, making results not completely reliable.

### Emissions

Polluting emissions can be used as a measure of stringency, in particular emissions in air and energy consumption. Scholars have considered the emission indicators in different ways; for some, a high number of emissions indicate that environmental regulation is too permissive; for others it demonstrates the tendency to a stricter environmental regulation. In particular, the literature on emissions as measure of stringency is divided in two branches: the first one considers regulation as exogenous, where environmental rules are given by external sources, such as a central government or an international organization. The second one considers emissions as a factor of production as any other in the firm's profit maximization dynamic.

With regard to the first case, some used the emissions' reduction as indicator for stringency (Smarzynska and Wei 2001; Gullop and Roberts 1983). For the second one, emissions produced by firms are intended as a factor of production like any other, with the consequence that environmental stringency can be assessed though the amount of emissions that the firm decides to produce. This approach is based on the neoclassical assumption of the profit-maximizing firms: a firm that decides to maximize its profits would use its factors of production until the marginal revenue of the product equals its price (Levinson 2001).

Currently there is no global instrument that collects emissions for each country˙ Datasets on emissions are available only for Europe and the United States. In Europe the *European Pollutant Release and Transfer Register* (E-PRTR), which replaced the *European Pollution Emissions Register* (EPER), collected data on emissions into air and water from 2007 to 2014. In the United States, the US Environmental Protection Agency listed a new dataset, the *Trade and Environmental Assessment Model* (TEAM). TEAM considers emission produced as factor inputs in the production process, in that it combines these data with the industry sector, location, and trade agreements, being therefore able to assess the effect that a certain trade agreement could have on pollution discharge (Creason et al. 2005). Such database is available for three different periods, 1997, 2002, and 2007, only for the United States.

## Indexes

Through the inclusion of individual indicators into a unique instrument, indexes can tackle many of the multidimensionality and complexity issues (Smarzynska and Wei 2001).

The first complete index is the one that Dasgupta et al. (2001) prepared in the occasion of the *United Nations Rio Conference on Environment and Development* on the basis of national reports prepared by public authorities. Public officials and NGO representatives had to answer questions regarding environmental awareness, the scope of environmental policies and regulations, the presence of control mechanism, and the implementation of environmental regulation. The analysis included four media (water, soil, air, biodiversity). However, such index was composed only for 1 year and afterward extended by Eliste and Fredriksson (2002) for the agricultural sector.

Another well-known index is the CLIMI (*Climate Laws, Institutions and Measures Index*) prepared for the European Bank for Reconstruction and Development (EBRD) in 2010 on the basis of the UN country reports and the UNFCCC (United Nations Framework Convention on Climate Change) reports. The CLIMI includes three dimensions: international cooperation, domestic climate framework, and fiscal or regulatory measures. The disadvantage of this index is that it considers only rules designed to address climate change and that it was drafted only for the year 2010 (Surminski and Williamson 2012).

The World Economic Forum used another method; businessmen and firm directors had to rate the regulatory stringency of the country in which the firm operates and to give them a grade from one to seven, where one represents very lax regulation and seven is the strictest environmental stringency. The result, called *Environmental Sustainability Index* (ESI), is complete, although it has the intrinsic shortcoming of being based on subjective perceptions and not on objective elements. The ESI has been used by Esty and Porter (2002) to build the *Environmental Regulatory Regime Index* (ERRI), together with the data on the Global Competitiveness Report 2001–2002 annual survey on business and government leaders (Esty and Porter 2005, p. 86).

EPI (Environmental Performance Index) measures environmental performance of countries in two main fields, human health and protection of the ecosystem. It is constructed through the use of 20 indicators reflecting national-level environmental data. The EPI represents a good source for panel data studies, since it examines more than 200 countries over 10 years. The benchmark for score application is the reaching of a “proximity target,” which corresponds to objectives fixed by international or national regulations, guaranteeing therefore a de facto assessment. Finally, the OECD has recently created a new index on environmental stringency, the Environmental Policy Stringency Index (EPS), which deserves a separate section for its peculiarities.

## Cross-References

▶ Environmental Policy: Choice

## References

Albrizio S, Koźluk T, Zipperer V (2014) Empirical evidence on the effects of environmental policy stringency on productivity growth. OECD Economics Department working papers, vol 1179. OECD Publishing, Paris

Becker RA, Henderson JV (2000) Effects of air quality regulations on polluting industries. J Polit Econ 108:379–421

Botta E, Kozluk T (2014) Measuring environmental policy stringency in OECD countries: a composite index approach, OECD economic department working papers, n. 1177

Brunel C, Levinson A (2013) Measuring environmental regulatory stringency. OECD trade and environment working papers, 2013/05. OECD publishing, Paris

Brunnermeier SB, Cohen MA (2003) Determinants of environmental innovation in US manufacturing industries. J Environ Econ Manag 45:278–293

Carraro C, De Cian E, Nicita L, Massetti E, Verdolini E (2010) Environmental policy and technical change: a survey. Int Rev Environ Resour Econ 4:163–219

Creason J, Fisher M, Morin I, Stone SF (2005) Comparison of the environmental impacts of trade and domestic distortions in the United States. Environmental economics working paper series, EPA, Washington DC, USA

Dasgupta S, Wheeler D, Roy S, Mody A (2001) Environmental regulation and development: a cross-country empirical analysis. Oxf Dev Stud 29:173–187

De Serres A, Murtin F, Nicoletti G (2010) A framework for assessing green growth policies. OECD Economic Department working papers, vol 774. OECD publishing, Paris

Eliste P, Fredriksson PG (2002) Environmental regulations, transfers and trade: theory and evidence. J Environ Econ Manag 43:234–250

Esty, Daniel, and Michael E. Porter (2002) Ranking national environmental regulation and performance: a leading indicator of future competitiveness? *In The Global Competitiveness Report 2001–2002*, by Michael E. Porter, Jeffrey D. Sachs, Peter K. Cornelius, John W. McArthur, and Klaus Schwab, 78–101. New York: Oxford University Press

Esty D, Porter ME (2005) Ranking national environmental regulation and performance: a leading indicator of future competitiveness? Available at http://www.hbs.edu/faculty/Publication%20Files/GCR_20012002_Environment_5d282a24-bb10-4a9a-88bd-6ee05e8c6678.pdf

Gallaher MP, Murray BC, Nicholson RL, Ross MT (2006) Redesign of the pollution abatement costs and expenditures (PACE) survey: findings and recommendations from the pretest and follow up visits. Final report for the US Environmental Protection agency, EPA, Washington DC, USA

Gray W (1997) Manufacturing plant location: does state pollution regulation matter? NBER working papers, vol 5880. National Bureau of Economic Research, Cambridge, MA

Gullop F, Roberts J (1983) Environmental regulations and productivity growth: the case of fossil-fueled electric power generation. J Polit Econ 91:654–674

Henderson JV (1996) Effects of air quality regulation. Am Econ Rev 33:789–813

Heutel G (2010) Plant vintages, grandfathering and environmental policy. Econ Fac Publ 14:1–59

Johnstone N, Hascic I, Kalmova M (2010) Environmental policy characteristics and technological innovation. Econ Polit 27:277–302

Keller W, Levinson A (2002) Pollution abatement costs and foreign direct investments inflows to U.S. states. Rev Econ Stat 84(4):691–703

Kozluk T (2014a) Measuring environmental policy stringency in OECD countries: a composite index approach. OECD economic department working papers, vol 1177. OECD Publishing, Paris, p 35

Kozluk T (2014b) The indicators of the economic burdens of environmental policy design: results from the OECD questionnaire. OECD Economics Department working papers, vol 1178. OECD Publishing, Paris

Levinson A (1999) An industry-adjusted index of state environmental compliance costs. NBER chapters behavioral and distributional effects of environmental policy. Cambridge, MA pp 131–158

Levinson A (2001) An Industry adjusted index of state environmental compliance costs. In: Behavioural and distributional effects of environmental policy. Universitty of Chicago Press, Chicago

Levinson A, Taylor MS (2008) Unmasking the pollution haven effect. Int Econ Rev 49:223

Malatu A (2008) Weighting the relative importance of environmental regulation for industry location. University of Manchester, Economics Discussion paper, EDP-0803

McConnell VD, Schwab RM (1990) The impact of environmental regulation on industry location decisions: the motor vehicle industry. Land Econ 66:67–81

Pearce D, Palmer C (2001) Public and private spending for environmental protection: a cross-country policy analysis. Fisc Stud 22:403–456

Sauter C (2014) How should we measure environmental policy stringency? IRENE working paper 14–01. Institute of Economic Research, University of Neuchatel, pp 4, 5

Smarzynska BJ, Wei SJ (2001) Pollution havens and foreign direct investment: dirty secret or popular myth? In: The BE journal of economic analysis and policy. Berkeley Electronic Press, 3(2):pp 1–34

Surminski S, Williamson A (2012) Policy indexes. What do they tell us and what are their applications? The case of climate policy and business planning in emerging markets. Working paper, vol 101. Centre for Climate Change Economics and Policy

Wing-Hung L, Zhan X, Liu N (2011) Understanding styles of corporate compliance with environmental regulation: towards a multidimensional conceptual framework. Paper prepared for the 11th national public management research conference Syracuse University

Xing Y, Kolstad C (2002) Do lax environmental regulations attract foreign investment? Environ Resour Econ 21:1–22

# Media

Peter T. Leeson and Joshua Pierson
Department of Economics, George Mason
University, Fairfax, VA, USA

**Abstract**

Citizens can use media to solve political agency problems. To be useful for this purpose, however, media must be free. Freer media are strongly associated with superior political-economic outcomes and may be especially important to fostering political-economic improvement in the developing world.

## Synonyms

Mass media

## Definition

Media refer to means of mass communication such as the Internet, television, radio, and newspapers.

## Media as a Mechanism for Controlling Government

Citizens in democratic regimes face a principal-agent problem with respect to their elected governors. While citizens empower political officials to wield government's authority in citizens' interests, left unchecked, political officials are tempted to use that authority for personal benefit. Democratic elections provide a means by which citizens may hold elected officials accountable for their uses of government authority. However, citizens' ability to use the voting both for this purpose depends crucially on the extent of their knowledge about political actors' behavior and their ability to coordinate responses to such behavior.

Media can improve democracy's ability to help citizens address the principal-agent problem they face with respect to elected officials (Besley and Burgess 2001, 2002; Coyne and Leeson 2004, 2009a, 2009b, Leeson and Coyne 2007). By reporting on political actors' behavior, media inform citizens about the activities of political actors that are relevant for citizens' evaluation of such actors as stewards of citizens' interests. Moreover, by providing such information to large numbers of citizens, media can help coordinate citizens' responses to what they learn about political actors' behavior, rewarding faithful stewards of their interests through, for example, reelection and punishing bad stewards through popular deposition and/or refusal to reelect them.

Media can also help citizens address political agency problems through the foregoing channel by influencing who seeks political office. Where would-be political officials know that the private benefits of holding political office are low because of media-provided information and media-facilitated citizen coordination, individuals who desire political power to further their own interests rather than citizens' are less likely to seek elected office.

## Media Freedom and Government Control of Media

The extent to which media can assist citizens in addressing the principal-agent problem they face with respect to political officials depends on media's freedom. Media freedom (or independence) refers to the extent to which government can directly or indirectly control the content of media-provided information reaching citizens. Where media freedom is higher, government's ability to influence the content of media-provided information is weaker and vice versa.

Government control of media can take many forms (Leeson and Coyne 2005). The most direct form is state ownership of media outlets. For example, all North Korean media outlets are controlled by the Korean Workers' Party or other appendages of the North Korean government. Elsewhere, media outlets are not owned by government, but are nevertheless owned by powerful people in government, creating a similar situation. Italy under the prime ministership of Silvio Berlusconi, who was also an Italian media mogul, is one well-known example of this phenomenon.

Government may also control media indirectly through ownership of media infrastructure. For instance, Romania's only newsprint mill was state owned for years following the end of Romanian communism. Similarly, the Associated Press of Pakistan is owned by the Pakistani government. In other countries, government exercises indirect control over media outlets whose financial positions depend on state-supplied income, such as revenue from government advertising.

Another important source of indirect government control of media is regulation of the media industry. Many governments require licenses for newspapers, television stations, and even journalists to operate and may use this power to restrict entry into the media industry to individuals who are friendly to the government and/or use the threat of license revocation to silence media critics.

Where government exerts significant influence over media and thus media are unfree, media's usefulness as a mechanism for assisting citizens to solve the agency problems they face with respect to their political officials is seriously impaired. Rather than monitoring political actors' behavior and reporting accurately on their uses of authority, media are likely to avoid furnishing citizens with such information or, worse still, furnish citizens with misleading information that benefit those in power. Uninformed or misinformed citizens find it difficult to use media-provided information to hold political actors accountable or to coordinate appropriate responses to such actors' behavior. Media's ability to effectively control government thus hinges critically on its freedom from government.

## Media's Influence on Politics

A strong link exists between media and citizen knowledge. Citizens who are exposed to more media coverage of their local politics and who live in regions where media are freer are more politically knowledgeable than citizens who enjoy less media coverage of their local politics and who live in regions where media are less free. For example, in Eastern Europe, citizens who live in countries where media are freer are more likely to correctly answer basic questions about political representation in the EU, and in the United States, citizens who are exposed to more media coverage of their local politics are more likely to know their congressman's name (Leeson 2008; Snyder and Stromberg 2010).

A strong link also exists between media and political-economic outcomes. Across countries, freer media are associated with higher voter turnout and other forms of political participation (Leeson 2008). Freer media are also associated with higher income, more democracy, more education, and more market-oriented economic policies (Djankov et al. 2003). In the United States, congressmen whose districts receive better media coverage are more likely to vote in a manner consistent with their party's line, stand witness before congressional committees, and procure more spending for their districts (Snyder and Stromberg 2010).

## Media's Role in Developing/Transition Economies

In developing and transition economies, media freedom is especially critical for political-economic development. Here, the problem of dysfunctional and corrupt government is pronounced, rendering media as a mechanism for controlling such malfeasance – a mechanism that, as indicated above, requires media independence – of particular importance.

Peru provides a striking example of how even a small amount of media independence can have a large effect on political-economic outcomes in the developing world (McMillan and Zoido 2004).

Despite the Fujimori government's bribe-secured control of all major media outlets in the country in the late 1990s, a small independent station that had not been bribed, Channel N, managed to acquire a recording of a high-ranking government official bribing an opposition politician to switch parties. Channel N repeatedly broadcast the video, and following suit, other channels began doing so too, eventually generating a popular backlash against the Fujimori government and the downfall of the corrupt Fujimori administration. In this case, the existence of only a single free media outlet proved critical to exposing political malfeasance and catalyzing the removal of a self-serving government.

In Russia, in contrast, a dearth of media freedom has prevented media from controlling government. Under the Soviet regime, government completely controlled Russian media, which served as little more than tools for government propaganda. During glasnost and perestroika, laws guaranteeing media independence were passed, and Russia's media appeared to become significantly freer. Ties between Russian media and political elites were never completely severed, however, and Russian media independence suffered major blows during the economic downturn of the early 1990s.

During this period, Russian media circulation and revenue plummeted. In consequence, many media outlets became dependent on state subsidies. Today, the Russian government commonly interferes with freedom of the press (on an important exception, see Enikolopov et al. 2011). Owners of media outlets that are critical of the government are threatened with jail time or forced to sell, and journalists who are critical of the state have been intimidated and even killed (Zassoursky 2004). Government's influence on media in Russia has contributed to Russian political actors' ability to wield public office for private gain.

## References

Besley T, Burgess R (2001) Political agency, government responsiveness and the role of media. Eur Econ Rev 45:629–640

Besley T, Burgess R (2002) The political economy of government responsiveness: theory and evidence from India. Q J Econ 117:1415–1451

Coyne CJ, Leeson PT (2004) Read all about It! Understanding the role of media in economic development. Kyklos 57:21–44

Coyne CJ, Leeson PT (2009a) Media, development, and institutional change. Edward Elgar, Northampton

Coyne CJ, Leeson PT (2009b) Media as a mechanism of institutional change and reinforcement. Kyklos 62:1–14

Djankov S, McLiesh C, Nenova T, Shleifer A (2003) Who owns the media? J Law Econ 46:341–382

Enikolopov R, Petrova M, Zhuravskaya E (2011) Media and political persuasion: evidence from Russia. Am Econ Rev 101:3253–3285

Leeson PT (2008) Media freedom, political knowledge, and participation. J Econ Perspect 22:155–169

Leeson PT, Coyne CJ (2005) Manipulating the media. Inst Econ Dev 1:67–92

Leeson PT, Coyne CJ (2007) The reformers' dilemma: media, policy ownership, and reform. Eur J Law Econ 23:237–250

McMillan J, Zoido P (2004) How to subvert democracy: Montesinos in Peru. J Econ Perspect 18:69–92

Snyder JM, Stromberg D (2010) Press coverage and political accountability. J Polit Econ 118:355–408

Zassoursky I (2004) Media and power in post-Soviet Russia. M.E. Sharpe, New York

## Medical Experimentation

▶ Human Experimentation

## Medical Liability

Ben C. J. van Velthoven and
Peter W. van Wijck
Leiden Law School, Leiden University,
Leiden, The Netherlands

### Abstract

This article concerns the preventive effects of medical liability. On theoretical grounds it is argued that medical liability does not necessarily lead to a socially optimal level of precaution, because the incentives are distorted in various ways. Since the 1970s, US states have enacted a variety of reforms in their tort systems. This variation has provided highly useful data for empirical studies of medical liability issues. For one thing, it has become clear that only some 2% of the patients with negligent injuries gets compensation. The empirical evidence nevertheless suggests that medical liability pressure does affect the behavior of healthcare providers to some degree. It has a negative effect on the supply of services, and it encourages the ordering of extra diagnostic tests. At the margin, medical liability law does seem to have some social benefits that offset reasonable estimates of overhead and defensive medicine costs.

## Definition

Medical liability essentially is tort law applied to healthcare providers. If negligent behavior of a healthcare provider causes harm to a patient, the healthcare provider has to pay damages to the patient. In this way, medical liability may lead to compensation of harm. Medical liability may also influence incentives to take care and consequently influence the probability and the size of harm. In the Law and Economics literature, the focus is on this preventive function.

## Introduction

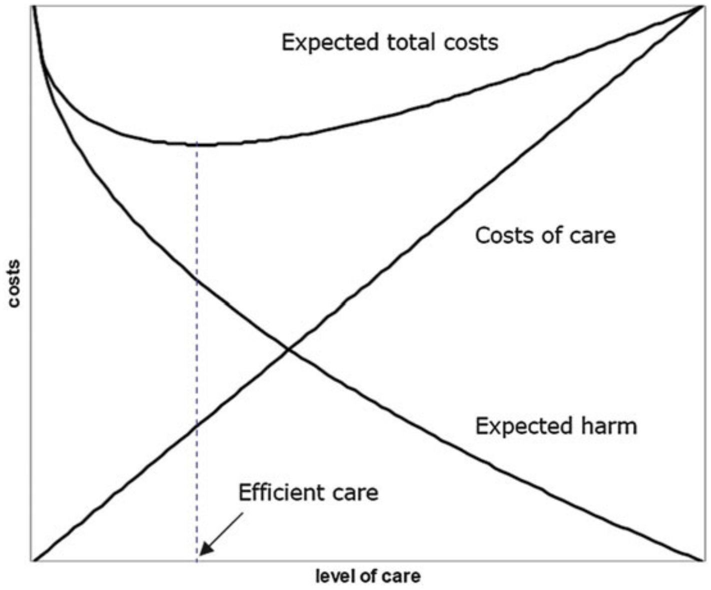
Medical treatment is supposed to improve the patient's health. But there is no guarantee. The patient's condition, bad luck, and medical errors may stand in the way of recovery. As adverse events occur within a market-type relationship, physicians and patients could write a complete contract to lay down the mutual understandings with respect to the specifics of the treatment and set the price of the contract accordingly. But this manner of controlling the level of care fails because of asymmetric information (Arlen 2013). Patients just do not know whether a physician delivers medically appropriate care. Moreover, individual bargaining about the level of care and the corresponding price brings along huge transaction costs and is even totally impossible in the case of emergencies. Fortunately, society has several other institutions available for the control of medical care. The government can centralize control by installing a regulatory body to enforce safety standards. State licensing, disciplinary boards, and hospital credential committees may also motivate physicians to act with proper care. Here we focus on medical liability (see also Van Velthoven and Van Wijck 2012).

Medical liability has received much attention in the Law and Economics literature in the past decades. The main reason is that since the 1970s, the USA has experienced three medical malpractice crises, periods characterized by significant increases in the premiums and contractions in the supply of malpractice insurance. In response to these crises, US states have enacted a variety of reforms in their tort systems. As a result, the USA has seen a considerable variation, across time and space, in the pressure of the medical liability system on health care providers. That variation has provided highly useful data for empirical studies into the actual working of the tort law system. For the same reason, most of those studies focus on the USA.

## The Standard Tort Model Applied to Medical Liability

In the economic analysis of tort law, a fundamental distinction is made between unilateral and bilateral accidents (Shavell 2004). Medical injuries can generally be taken to be *unilateral accidents*. A physician who wants to reduce the probability and severity of medical injury can increase the number of visits provided to his patient, perform additional diagnostic tests, refer the patient to a specialist, opt for more or less invasive procedures, and/or take more care in performing surgery. The patient is usually unable to influence expected harm from a medical injury.

Figure 1 presents the standard tort model for a unilateral accident case (Miceli 2004, pp. 42–45). As the (potential) injurer raises the level of care by taking additional precautionary measures, his *costs of care* increase. But at the same time, there is a reduction in the *expected harm* for the



**Medical Liability, Fig. 1** Efficient care

(potential) victim, as additional care may reduce the probability and/or the severity of accident losses. The social optimum is obtained if the expected total costs are at a minimum. The socially optimal level of precaution is frequently referred to as the *efficient level of care*.

Medical liability law quite universally holds an injurer liable for accident losses that are attributable to *negligence*. The negligence rule presupposes a norm of *due care*, specified by statutory law or jurisprudence, for the precautionary measures that the injurer should take at a minimum. If the injurer's level of care falls short of this minimum, the injurer is negligent and will be held liable for accident losses. On the other hand, if the injurer's level of care equals or exceeds the due care norm, accident losses will remain with the victim. In this way, the negligence rule creates a discontinuity in the injurer's expected costs at the level of due care, as shown by the fat curve in Fig. 2. The injurer minimizes his costs by just taking precautionary measures in conformity with the due care norm.

Whether the injurer will act in a socially optimal manner then depends on the proper choice of the due care norm. He will take socially optimal precaution if the level of due care coincides with the efficient level of care, as in Fig. 2. If the due care norm is set below (above) the efficient level of care, the personal incentives will generally lead the injurer to behave in a suboptimal manner by taking too little (too much) precaution.

Apart from the level of care chosen by the (potential) injurer when engaging in a certain activity, total costs for society also depend on his *number of activities.* Under negligence, when the injurer conforms to the due care norm, he cannot be held liable for any accident losses. Consequently, a non-negligent injurer is not confronted with the full social costs of his activity. This negative externality may lead the injurer to undertake too many activities from a social point of view.

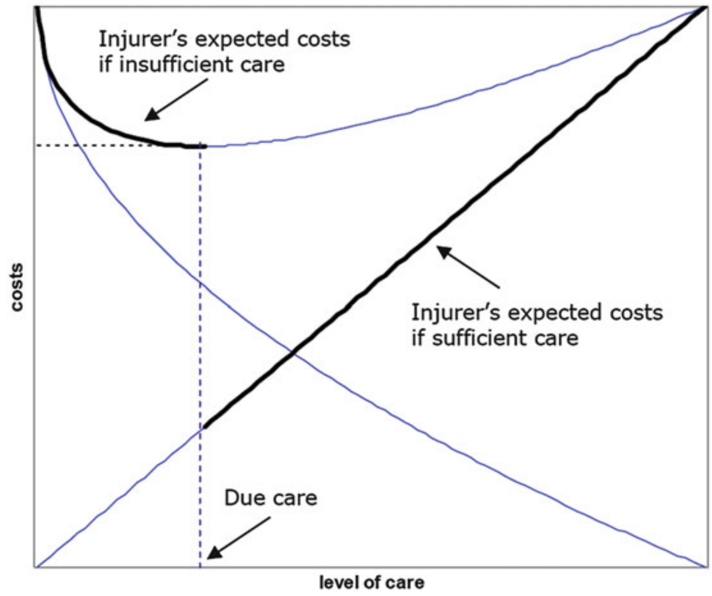
## Problems in Applying the Standard Tort Model to Medical Liability

In practice, a number of specific characteristics of medical care pose serious problems to a straightforward application of the standard tort model.

### Uncertainty About Due Care

The standard tort model suggests that injurers can be induced to provide socially optimal care if law

**Medical Liability, Fig. 2** Efficient care under the negligence rule



sets due care at the efficient level. This simple rule, however, is not easily met in the health care sector:

- It is difficult to determine the efficient level of care, even for specialists in the field. Instead, the courts generally evaluate the conduct of physicians in terms of customary standards of practice within the medical profession or a significant minority of such professionals (Weiler et al. 1993, p. 8; Danzon 2000, p. 1343; but see Peters (2002) for a somewhat different reading).
- Physicians have no exact insight into the due care norm that will be applied by the courts when confronted with a claim.
- In each specific case, the court has to decide whether the physician has been careful enough. Even if the due care norm is right, the court may err in its decision, as the information presented to the court will generally be incomplete and subject to mistakes in interpretation. Consequently, physicians do not know exactly how careful they have to be in order to escape liability.

Calfee and Craswell (1984) have studied the consequences. If there is uncertainty about the due care norm and its application in court, there is a chance that a physician who has taken sufficient care may still be held liable for damages. The physician can try to reduce that chance by *overcomplying*, that is, by raising his level of care beyond due care.

## Ethical Norms

Most physicians swear (some variant of) the Hippocratic oath to use their best ability and judgment in treating their patients. This may affect the ethical values in the profession in such a profound manner that all other (financial) considerations are rendered more or less futile. If so, the result would be overcompliance, that is, delivering more care than strictly necessary.

## Who Bears the Costs of Care?

The standard tort model starts from the premise that the costs of care are borne by the injurer. This is, however, not self-evident in the health care sector, as physicians are generally paid on a fee-for-service basis. Moreover, most patients carry a health insurance policy. When a physician decides to raise the level of care, he will not have much trouble in charging the additional costs to his patient, who will forward the bill to his insurance company.

Of course, this financial incentive to provide too much care will be mitigated if healthcare services are financed in a different way, which hinders or prevents the passing on of additional care costs. One can think of: pure salary payment, capitation payment, or managed care plans (Glied 2000).

### Litigation Problems

The standard tort model takes it for granted that the injurer pays damages, once he is found negligent. In general, however, this payment will not be made spontaneously. First, the victim has to decide whether it is in his interest to file a claim. The patient may be unaware of the negligence of his physician, the decision of the court may be too uncertain, the litigation costs may be too high, or the financial means of the victim may be insufficient. Second, if the claim is below a threshold of, say, $250,000, the patient may not be able to find an attorney willing to accept his case on a contingent fee basis (Shepherd 2014). Third, even if a claim is filed, parties may decide to settle the case in order to save on litigation costs. If the case is settled, damages will not amount to a full compensation.

The implication of all this is that expected damage payments are smaller than the expected harm of the injurer's behavior. The result is a tendency toward insufficient care.

### Liability Insurance

Most, if not all physicians, carry a liability insurance policy (GAO 2003, p. 6). By shifting the burden of the damages, liability insurance lowers the injurer's costs of insufficient care (Zeiler et al. 2007).

In theory, experience rating can redress this tendency. Experience rating refers to a variety of schemes that see to it that liability insurance premiums reflect each insured's expected loss (Danzon 2000, p. 1160). This is mostly done by varying premiums with past claims or loss experience. Shavell (1982) has shown that insurance need not interfere with the incentive effects of liability if premiums are perfectly experience rated. But experience rating is more easily said than done in the case of medical care. Medical malpractice claims occur too infrequently to give insurance companies enough information to reliably set premiums in accordance with individual physician's care levels.

Medical liability insurance, however, does not completely eliminate incentives to take care. Malpractice may affect the physician severely, even if he is fully insured against the financial consequences, as claims also bring along other kinds of costs. The defense takes quite a lot of precious time, the experience is rather unpleasant, and it may cause serious reputational harm (Weiler et al. 1993, p. 126).

### Patient Safety Investments

In the standard tort model, the individual physician makes a more or less informed decision on how to treat each particular patient. But that model does not fully capture the causes of medical negligence (Arlen 2013). Many medical injuries occur accidentally when physicians unknowingly err in the diagnosis and the selection of treatment or when the hospital equipment and staff fail in delivering the treatment. The probability of such errors can be reduced by investing in expertise, healthcare technology, and personnel. But such patient safety investments generally fall outside the factors courts consider in determining whether a physician has been negligent. The incentives for the physician, and the hospital, to invest in patient safety are therefore suboptimal.

### Conclusion

Medical liability does not necessarily lead to a socially optimal level of precaution, because the incentives are distorted in various ways. For one thing, physicians generally do not bear the full accident losses of insufficient care. This distortion may act as an invitation to physicians to take less care than legally required. Still, the nonfinancial consequences of liability (time, hassle, reputation loss) and ethical considerations may provide some counterweight. Other distortions provide incentives to act on the safe side of the due care norm. Physicians generally do not bear the (full) costs of care due to specific methods of financing in the healthcare sector. And there is uncertainty about the due care norm and its application by the courts. On balance, there might be a bias toward excessive care.

## Defensive Medicine

In the USA, the conviction has taken root among physicians and their liability insurers that the medical liability system has gone wrong. It is argued that the pressure has evolved to such a level that it has given rise to *defensive medicine*. The most common definition (OTA 1994, p. 21) reads: "Defensive medicine occurs when doctors order tests, procedures, or visits, or avoid high-risk patients or procedures, primarily (but not necessarily solely) to reduce their exposure to malpractice liability." According to this definition, defensive medicine can take two forms. *Positive* defensive medicine involves supplying care that is not cost effective, unproductive, or even harmful. *Negative* defensive medicine involves declining patients that might benefit from care. It also includes physicians deciding to exit the profession altogether.

Our discussion of the standard tort model demonstrates that positive defensive medicine will not necessarily be found in practice, as liability pressure on the level of care is working in two opposite directions. Thus, the question whether physicians take excessive care is really an *empirical* question. Second, if malpractice pressure does produce a bias toward excessive care, it is excessive in comparison to the *due care* norm. But it is not at all certain that the due care norm has been set equal by law to the efficient level of care. That leaves the possibility, even if empirical research finds proof of excessive care, that level of care still falls short of the socially optimal amount (Sloan and Shadle 2009, p. 481).

The concept of negative defensive medicine is related to the number of activities, referred to earlier. If a physician takes at least due care, he will not be liable for any accident losses, which might give him the incentive to accept too many patients from a social point of view and/or to stay too long in the profession. On the other hand, the simple fear of malpractice claims, even if unwarranted, and the corresponding threat of time and reputation loss may work in the opposite direction. Hence, the existence and scope of negative defensive medicine is, once again, an *empirical* question. Moreover, note that the interpretation of the findings may change if physicians exercise insufficient care. Then, malpractice law helping patients to file claims and to obtain damage payments may give negligent physicians a good reason to revise their conduct, not only by raising the level of care but also by accepting fewer patients or by early retiring. Such a behavioral response might be very welcome from a social point of view.

What makes the interpretation of the findings even more complicated is the interaction between the defensive medicine that may follow from medical liability pressure and the *offensive medicine* that is induced by physicians' financial incentives (Avraham and Schanzenbach 2015). When physicians have the discretion to choose among different treatment regimens and health insurance adequately covers their patients' medical costs, physicians may be tempted to opt for the more invasive procedures, which in general will be the more remunerative ones. But more invasive procedures are also riskier. Hence, medical liability pressure may counteract the tendency toward offensive medicine.

## Medical Liability Litigation

This section surveys the empirical evidence concerning medical liability litigation. The different layers of the dispute pyramid (Galanter 1996) are discussed one by one.

Three large-scale surveys of medical records of hospitalized patients in the USA have investigated the incidence of injury due to negligent medical care. In the most recent survey in Utah and Colorado in 1992 (Studdert et al. 2000), it was found that 2.9% of all hospitalized patients had an adverse event that was related to medical care. Some 0.8% of the patients suffered a negligent injury, where negligence was defined as treatment that failed to meet the standard of the average medical practitioner. No attempt was made to define negligence by weighing marginal costs and benefits of additional precautions. So, the resulting count of negligent injuries does not necessarily correspond to inefficient injuries.

The second layer of the dispute pyramid discloses how many of the injury victims take steps to obtain compensation. In the Utah and Colorado study, only 3% of the patients who were identified as having sustained a negligent injury filed a malpractice claim. But there was also a significant number of "false positives." Aggregate data from insurers' records pointed out that a substantial number of malpractice claims do not correspond to an identifiable injury due to negligent medical behavior. Of course, all these plaintiffs may still have filed the claims in good faith, from a state of imperfect information, leaving it to the tort system to separate the rightful claims from the non-deserving ones.

The third layer of the dispute pyramid discloses how filed claims fare in the tort system. In a large study of malpractice claims closed between 1984 and 2004, Studdert et al. (2006) found that 61% of claims could be associated with injury due to medical error, while 39% of the claims had no merit. Only 15% of all the claims were resolved by trial verdict; the rest was settled in the "shadow of the law" or dropped. Most of the claims involving injuries due to medical error (73%) received compensation; most claims not involving medical error (72%) did not receive compensation. Moreover, when claims involving error were compensated, payments were significantly higher on average than were payments for non-error claims.

With respect to the payment amounts, two observations are in place. First, compensation in most cases falls short of plaintiff's losses, especially for more serious injuries (Sloan and Chepke 2008). Second, the costs of administering the tort system (legal expenses, overhead costs) are considerable. According to calculations by Mello et al. (2010), it costs US society overall more than $1.70 to deliver $1 of net compensation.

The tort litigation system is not perfect, then. It sometimes makes physicians – or their insurers – pay damages for non-negligent care. But the system is clearly not a random lottery (see also Eisenberg 2013). Negligent injuries are at least ten times as likely to end up in compensatory payments as non-negligent injuries. As a result, there is a strong association between the numbers of adverse patient safety events in hospitals and paid medical malpractice claims (Black et al. 2017).

More disturbing for the proper working of the system is the high rate of "false negatives." From the figures above, it follows that just some 2% of the patients with negligent injuries gets compensation, mainly because a large fraction of valid claims is not filed, but also because not all valid claims that are filed get honored. And even that 2% is quite likely a serious overestimation, as an observational study of healthcare providers has shown that hospital records may miss over 75% of serious medical errors (Andrews 2005). Combining the high rate of false negatives with the finding that compensation generally falls short of victims' losses suggests that the deterrent function of the system must be rather limited.

## Tort Reform

In the introduction, it was noted that the USA experienced three "crises" in the medical liability insurance market in the past decades. These were periods of deterioration in the financial health of carriers, followed by sharp increases in premiums and contractions in supply. This is not the place to delve deeply in the causes of these crises (cf. Danzon 2000; Sloan and Chepke 2008). But one factor can be singled out: the "long-tail" character of this line of insurance. Claims may be filed many years after an adverse event causes injury. And from there, it may take many more years before the insurance company finally knows how much compensation it has to pay. If, for whatever reason, there is a gradual rise in claim frequency and/or in average payments, for instance, because of pro-plaintiff adaptations in common law doctrines or because patients are becoming more assertive toward healthcare professionals, insurance companies will tend to lag behind. They will develop unexpected losses, and overreact in raising premiums and curtailing supply.

In response to the malpractice crises, most US states have adopted tort reform measures. The objective of these measures is to reduce the overall costs of medical liability. The extent and specifics of tort reform vary from state to state (cf. www.atra.org). Some reforms aim at a reduction of damage awards, other reforms make it more costly or difficult to file tort cases in the first place. The most commonly adopted tort reforms are: caps on non-economic damages, pretrial screening panels, contingency fee reform, joint-and-several liability reform, collateral source rule reform, periodic payment, and shorter statutes of limitation. The effects of these tort reforms on the frequency and the size of claims and on malpractice insurance premiums have been studied extensively. A detailed review by Mello and Kachalia (2016) concludes that there is no convincing evidence that any other reform than caps on non-economic damages has had a significant impact. As to damage caps: the weight of the evidence suggests that they reduce claims frequency, achieve substantial savings in average damage payments, and modestly constrain the growth of malpractice insurance premiums. So one would be tempted to conclude that at least this specific kind of tort reform can help to relieve malpractice pressure, if so desired. But even that conclusion is called into question. Zeiler and Hardcastle (2013) point out that thus far no one study has employed a consistently solid set of empirical research methods.

More recently, also other kinds of reform measures have been proposed, such as apology laws and disclosure programs, presuit notification periods, health courts, and safe harbors for adherence to evidence-based practice guidelines. As these reforms are relatively new in use, if at all, the empirical literature on their effects is as yet very small.

## Preventive Effects of Medical Liability

It is an empirical question whether medical liability leads physicians to take appropriate precautions or to engage in defensive medicine. In the literature, four main research lines can be distinguished.

The first line of research surveys physicians and asks their opinion on the role of malpractice pressure in clinical practice. These studies (e.g., Carrier et al. 2010) unequivocally point out that concerns about malpractice liability are pervasive among physicians. Indeed, in a survey of high-risk specialists, 93% of the interviewees reported practicing defensive medicine (Studdert et al. 2005). Yet, the results should be handled with caution. First, the relationship between perceived malpractice threat and objective liability risk is found to be very modest. Physicians systematically overestimate the risk that malpractice action will be brought against them. Second, the relationship between malpractice threat and clinical response is a self-reported one.

The second line is about the actual relevance of positive defensive medicine. How do treatment choices by physicians, and the health outcomes of their patients, respond to malpractice pressure? Much attention has gone to obstetrics, the field that has one of the highest levels of premiums, claim frequency, and damage payments. Typically, studies examine the impact of tort reform on cesarean section rates. Some studies have also looked at the impact on infant health at birth. Overall, the results are inconclusive. Thus far, there is no decisive evidence for positive defensive medicine in obstetrics (Eisenberg 2013). Some other studies focus on cardiac illness or take a look at broader sets of ailments or total healthcare expenditures. The results are mixed. As far as physicians are found to practice positive defensive medicine, the excessive care appears to be related to rather elementary diagnostic tests such as imaging, not to major surgical procedures. The overall picture is that the total effect on healthcare costs, if any, is rather small (Thomas et al. 2010).

The third line of research is on negative defensive medicine and analyzes how tort reform affects the supply of healthcare services. The evidence with respect to obstetrics is mixed. Other studies analyze the overall supply of physician services. Their results generally point out that higher malpractice pressure tends to diminish healthcare supply, be it the number of physicians, statewide or in local areas only, or their hours worked. That finding seems to be proof of negative defensive medicine. But note that the interpretation is not so obvious. A smaller supply of physicians in itself can be presumed to contribute negatively to social welfare, but there may also be offsetting effects if the quality of the physicians that stop or reduce their practice is below average. Indeed, Dubay et al. (2001) and Klick and Stratmann (2007) find no evidence that the changes in supply had negative health effects.

Finally, an interesting new line of research tries to assess the preventive impact of medical liability forces by drawing on variations in the negligence standard facing physicians. The majority of US states have over time moved from a due care norm based on the customary practices of local physicians to a national standard of care. The first empirical results (Frakes and Jena 2016) indicate that treatment quality improves when the clinical standards go up.

## Cost-Benefit Analysis

Empirical evidence suggests that medical liability pressure does affect the behavior of healthcare providers. It does seem to encourage the ordering of extra diagnostic tests, and it tends to reduce the supply of services. However, positive defensive medicine does not have a clear-cut effect on health. If the additional tests and procedures have any value, it is only a marginal one. Furthermore, changes in the supply of services do not affect health adversely. This suggests that the physicians that are driven out of business have a below average quality of performance. Hence, at the margin, medical liability law may have some social benefits (see also Zabinski and Black 2015).

These benefits must be weighed against the costs of the additional tests and procedures. The costs of administering malpractice claims also deserve attention. Both Danzon (2000) and Lakdawalla and Seabury (2012) have made a shot at a back-of-the-envelope calculation of the costs and benefits. They conclude that under quite general assumptions, the benefits of even a modest reduction in injury rates suffice to offset reasonable estimates of overhead and defensive medicine costs. This follows from the large social costs of medical injuries and the low rate of claims per negligent injury.

Yet, instructive as these calculations may be, they mainly have a heuristic value. First, a full cost-benefit evaluation of the medical liability system is impossible in the current state of affairs. Second, even if the marginal benefits of the current system do outweigh the costs, the search for improvements and alternatives is open (see, e.g., Sloan and Chepke 2008). It is argued that the impact of the medical liability system can be substantially improved by shifting liability from the individual physician to the medical entity involved (Arlen 2013) and by restructuring the financial incentives in the healthcare sector (Frakes 2015).

## Cross-References

- ▶ Litigation and Legal Expenses Insurance
- ▶ Mass Tort Litigation: Asbestos
- ▶ Medical Malpractice
- ▶ Strict Liability Versus Negligence
- ▶ Tort Damages

## References

Andrews L (2005) Studying medical error in situ: implications for malpractice law and policy. DePaul Law Rev 54:357–392

Arlen J (2013) Economic analysis of medical malpractice liability and its reform. In: Arlen J (ed) Research handbook on the economics of torts. Edward Elgar, Cheltenham, pp 33–68

Avraham R, Schanzenbach M (2015) The impact of tort reform on intensity of treatment: evidence from heart patients. J Health Econ 39:273–288

Black BS, Wagner AR, Zabinski Z (2017) The association between patient safety indicators and medical malpractice risk: evidence from Florida and Texas. Am J Health Econ 3:109–139

Calfee JE, Craswell R (1984) Some effects of uncertainty on compliance with legal standards. Virginia Law Rev 70:965–1003

Carrier ER et al (2010) Physicians' fears of malpractice lawsuits are not assuaged by tort reforms. Health Aff 29:1585–1592

Danzon PM (2000) Liability for medical malpractice. In: Culyer AJ, Newhouse JP (eds) Handbook of health economics, vol IB. North-Holland, Amsterdam, pp 1339–1404

Dubay L et al (2001) Medical malpractice liability and its effect on prenatal care utilization and infant health. J Health Econ 20:591–611

Eisenberg T (2013) The empirical effects of tort reform. In: Arlen J (ed) Research handbook on the economics of torts. Edward Elgar, Cheltenham, pp 513–550

Frakes MD (2015) The surprising relevance of medical malpractice law. Univ Chicago Law Rev 82: 317–391

Frakes M, Jena AB (2016) Does medical malpractice law improve health care quality? J Public Econ 143: 142–158

Galanter M (1996) Real world torts: an antidote to anecdote. Md Law Rev 55:1093–1160

GAO (2003) Medical malpractice insurance. Multiple factors have contributed to increased premium rates. US General Accounting Office, Washington, DC. Report GAO-03-702

Glied S (2000) Managed care. In: Culyer AJ, Newhouse JP (eds) Handbook of health economics, vol IA. North Holland, Amsterdam, pp 707–753

Klick J, Stratmann T (2007) Medical malpractice reform and physicians in high-risk specialties. J Leg Stud 36: S121–S142

Lakdawalla DN, Seabury SA (2012) The welfare effects of medical malpractice liability. Int Rev Law Econ 32:356–369

Mello MM, Kachalia A (2016) Medical malpractice: evidence on reform alternatives and claims involving elderly patients. A report prepared for the Medicare Payment Advisory Commission. MedPAC, Washington, DC

Mello MM et al (2010) National costs of the medical liability system. Health Aff 29:1569–1577

Miceli TJ (2004) The economic approach to law. Stanford University Press, Stanford

OTA (1994) Defensive medicine and medical malpractice. US Congress, Office of Technology Assessment, Washington, DC. Report OTA-H-602

Peters PG Jr (2002) The role of the jury in modern malpractice law. Iowa Law Rev 87:909–969

Shavell S (1982) On liability and insurance. Bell J Econ 13:120–132

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

Shepherd J (2014) Uncovering the silent victims of the American medical liability system. Vanderbilt Law Rev 67:151–195

Sloan FA, Chepke LM (2008) Medical malpractice. MIT Press, Cambridge, MA

Sloan FA, Shadle JH (2009) Is there empirical evidence for "defensive medicine"? A reassessment. J Health Econ 28:481–491

Studdert DM et al (2000) Negligent care and malpractice claiming behavior in Utah and Colorado. Med Care 38:250–260

Studdert DM et al (2005) Defensive medicine among high-risk specialist physicians in a volatile malpractice environment. JAMA 293:2609–2617

Studdert DM et al (2006) Claims, errors, and compensation payments in medical malpractice litigation. New Engl J Med 354:2024–2033

Thomas JW, Ziller EC, Thayer DA (2010) Low costs of defensive medicine, small savings from tort reform. Health Aff 29:1578–1584

Van Velthoven BCJ, Van Wijck PW (2012) Medical liability: do doctors care? Recht der Werkelijkheid 33(2):28–47

Weiler PC et al (1993) A measure of malpractice: medical injury, malpractice litigation and patient compensation. Harvard University Press, Cambridge, MA

Zabinski Z, Black BS (2015) The deterrent effect of tort law: evidence from medical malpractice reform. Northwestern University Law School, Chicago, IL. Law and Economics Research Paper No. 13-09

Zeiler K, Hardcastle L (2013) Do damages caps reduce medical malpractice insurance premiums? A systematic review of estimates and the methods used to produce them. In: Arlen J (ed) Research handbook on the economics of torts. Edward Elgar, Cheltenham, pp 551–587

Zeiler K et al (2007) Physicians' insurance limits and malpractice payments: evidence from Texas closed claims, 1990–2003. J Leg Stud 36:S9–S45

# Medical Malpractice

Veronica Grembi
Copenhagen Business School, Copenhagen, Denmark

## Abstract

MM first came to the attention of policy makers primarily in the USA where, from the 1970s, healthcare providers denounced problems in getting insurance for medical liability, pointing out to a *crisis* in the MM insurance market (Sage WM (2003) Understanding the first malpractice crisis of the 21th century. In: Gosfield AG, (ed) Health law handbook. West Group, St. Paul, pp 549–608). The crisis was allegedly grounded in an explosion of requests of compensations based on suffering iatrogenic injuries. Since then, MM problems have been identified with scarce availability of insurance coverage and/or its affordability, the withdrawal from the MM insurance of commercial insurers, the growth of MM public insurance or self-insurance solutions, the choice of no-fault

rather than negligence liability, the adoption of enterprise liability for hospitals, the concerns for defensive medicine, and the implementation of tort reforms so to decrease MM pressure (i.e., frequency of claims and the levels of their compensation) on healthcare practitioners. While the initial contributions to the topic are mainly based on the US healthcare and legal system experience, a growing attention to these problems has raised in the last decades also among European countries (Hospitals of the European Union (HOPE) (2004) Insurance and malpractice, final report. Brussels, www.hope.be; OECD (2006) Medical malpractice, insurance and coverage options, policy issues in insurance n.11; EC (European Commission, D.G. Sanco) (2006) Special eurobarometer medical errors).

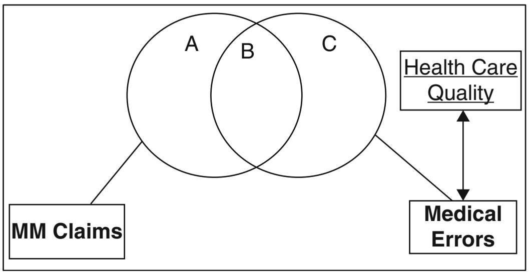
## Definition

Medical malpractice (MM) deals with two kinds of problems that healthcare providers, patients, and insurance companies have to deal with in different legal and healthcare systems: (1) medical errors without legal consequences and (2) medical errors with legal consequences. The contributions to this field aim at defining an efficient level of (1) and to efficiently deter and compensate (2). Empirical evidence on the actual trends of medical errors, medical liability claims, and the link of the former and the latter with both the trend of MM insurance premium and the treatment decisions of healthcare providers are not always unambiguous.

## Medical Malpractice

Starting from the 1970s, healthcare practitioners, lawyers, and insurance companies in the USA experienced so-called crises of MM. The characteristics of the first crisis, joint to those which followed during the 1980s and the beginning of the New Millennium, could be summarized as follows (Mello et al. 2003): first, an increase in MM claims not always justifiable by a similar increase in medical errors; second, an increase in MM premium for healthcare providers joint to the scarcity of available insurers in the market; and third, the tendency of healthcare practitioners to adopt behaviors which can minimize the probability to be targeted by a legal claim but not the probability of an error or bad quality medicine, also known as defensive medicine. All in all, there is an increase in legal, insurance, and healthcare expenditures, which have consequences for the quality of the delivered healthcare (Arlen 2013). Similar concerns and complaints have started to be common also among several European countries, notwithstanding their quite different healthcare and legal systems compared to the USA. The contribution of the economic analysis of law to this subject is at first a theoretical analysis of the role of legal rules as incentivizing mechanisms to achieve efficiency across different institutional settings. As such, the law and economics of MM provides a common ground of theoretical elements, which, in time, have been accompanied by the production and discussion of often controversial empirical evidence.

From a theoretical viewpoint, MM has to do with the definition of the *efficient* medical error as it has been shaped by the contributions of the economic analysis of tort law (Shavell 1987; Arlen 2013). According to this approach, the goal of legal rules is not to reduce the probability to have errors to zero, but to deter *inefficient* errors and compensate the innocent victim of an *inefficient* error. Suppose that the probability that an error $i$ takes place is equal to $p_i$ and that in case $i$ happens, victim $v$ will suffer damages equal to $D$ (both economic and noneconomic). The expected damages are equal to $p_iD$. The healthcare provider-potential injurer in this framework can decrease $p_i$ investing in precaution, which will cost her $c$, with $c$ increasing as $p$ decreases (i.e., $c(p)$). The efficient level of precaution minimizes the sum of the expected damages, $p_iD$, and the cost of precaution, $c$. For that optimal level of precaution, let us call it $x^*$, the level of medical errors allowed into the system is considered efficient and expected to be different from zero (i.e., the social optimum). We have tried to summarize this basic intuition in the diagram presented in Fig. 1. Areas B and C represent the



**Medical Malpractice, Fig. 1** Relationship between MM claims and medical errors

universe of medical errors within a given institutional setting. Consistently with what we have just stated, legal rules need to define the prerequisites to put an error in either B or C: in an ideal setting, if the legal system is properly designed, an error in C is efficient and an error in B is inefficient, and the dimension of B has to be efficient in a social welfare perspective.

In the same setting, a policy maker can set the boundaries between C and B using either a negligence/fault rule or a strict liability/no-fault rule. In the first case the stress is on the threshold of precaution below which the potential injurer is held liable. As in the case of a car accident liability, a speed threshold is set and if an accident occurs, you will be liable only if your speed was higher than the imposed limit. In the second case (i.e., No Fault) a key role is played by the causation link between the negative outcome and the injurer action. In other words you are responsible if you injure somebody in a car crash, but you are not accountable for injuries that your potential victim got before the crash. Both liability rules find a place in the realm of MM. Both can generate less than or more than efficient precaution in the real world.

The problem of transposing the basic liability model on medical "accidents" is that it is not always clear how precaution is related to expected outcomes. While in the case of a car accident, the sequence is clear: you speed, the car crashes, and a person that before the impact was sound is now injured. In the case of a medical accident, there might be a less clear sequence. A physician does not take the necessary precaution; she treats a sick patient, and the sick patient does not heal. Unfortunately it is not so simple. It could well be that the sick patient heals or that he/she heals but it takes a week more than expected. In other words, defining precaution in this context is not so straightforward. We could think about appropriate hygienic conditions and basic working environment, but then the real issue at stake is more related to the case of a misdiagnosis or to a mistreatment and to the elements causing the former rather than the latter. But this is a case-by-case call. Hence, the problem is how we define a case-by-case standard of care. For instance, the consensus on precaution levels (or treatment choices) can be high on routinely practiced interventions and low on more complicated procedures. Whenever the standard of care to be adopted is not accurately specified, the system might end up characterized by less or more than efficient standard of care. One consequence is the so-called defensive medicine, a modification in care decisions triggered by MM pressure (i.e., frequency of claims and the levels of their compensation), which can be positive or negative (Danzon 2000; Kessler 2011). Positive defensive medicine consists in the use of treatments or diagnostic tools that are not able to improve the quality of care delivered to patients, but apt to decrease the probability to be targeted by a legal claim. It is a form of supplied induced demand: if the patient has the same information that her physician has, she would have not chosen the recommended care. Negative defensive medicine coincides with forms of cream skimming of patients or procedure. Less risky patients are selected into treatment so to decrease the probability of negative outcomes, and needed risky treatments can be avoided due to the fear of the legal consequences.

A second challenge to the liability model is represented by the organization of the healthcare system. The healthcare organization matters as much as legal rules, to achieve or miss an efficient level of precaution. For instance, a physician could have different incentives to practice defensive medicine depending on whether she is employed by the hospital or she is an independent practitioner. Finally, the possibility to buy insurance for medical liability and the type of insurance can alter the structure of incentives produced by the legal rules. Insurance is available in a private/ commercial form, often not experience rated, or in a public form (However, the distortions generated by the insurance for medical liability should be attenuated by reputational concerns, supposedly more for private rather than public healthcare providers). Forms of public insurance can come together with hospital self-insurance within a public healthcare system, or through a specific public fund. Public coverage can solve some of the problems connected to the insurance market for MM, but they have the potential to generate new kind of problems, such as common pooling of individual risks among the covered healthcare providers, basically incentivizing under-precaution.

In other words the perceived costs/benefits of taking precaution are affected by the certainty of the standard of care, the organization of the healthcare system, and the availability and form of insurance for MM. In principle a fault system with a private MM insurance and no unanimity on the standard of care could trigger positive defensive medicine. Such consequence could be mitigated if the practitioner is not directly responsible, but as an employee of a hospital is covered by an enterprise liability. However, in this case less than efficient precaution could be held. For the same token, a no-fault liability system with private MM insurance could more easily generate negative defensive medicine behaviors. A public insurance scheme could cope with this problem, but again at the expenses of incentive to take precaution. In reality, MM liability comes together with other institutional elements able to affect the structure of incentives foreseen by the economic analysis of law, and this is why defining the best solution according to a social welfare perspective is a debated issue, which needs to be supported by sound empirical work.

Empirical evidence on MM can be grouped in two sets: (1) evidence on the incidence of medical errors, useful to assess the dimension of the iatrogenic injury problem, the relationship between actual errors and claims, and the inner causes of medical errors (Weiler et al. 1993; Nys 2009), and (2) evidence on the effectiveness of policies directly or indirectly oriented to decrease MM pressure on the healthcare providers as well as changing the structure of costs and benefits of potential claims (i.e., tort reforms) on (a) liability measures such as the frequency of MM claims and medical liability insurance and (b) care-related measures as defensive medicine, the quality of care, or the supply of physicians (Kachalia and Mello 2011; Kessler 2011).

Overall, the empirical assessment of MM is not an easy task. The number of claims filed every year against hospitals and medical practitioners might not necessarily stem from negligent mishaps, while many negligent behaviors and their outcomes are not actually prosecuted (Weiler et al. 1993). Defining for empirical purposes, a medical error or, more properly, an injury due to inacceptable medical negligence is not a straightforward task either: as stated it requires an implicit assumption on the expectations of the outcome of a specific treatment or procedure conditional to many variables describing the conditions of the patient (Danzon 2000). Expectations on a normal distribution of outcomes are not always set before running empirical investigations. The US studies are usually the benchmarks in this field. The first surveys had been run during and in the aftermath of the first malpractice crisis. The 1974 California Study is worth mentioning even though the most famous work is definitely the 1984 Harvard Study, dealing with New York Hospitals data (Weiler et al. 1993). Later on similar initiatives have been undertaken in other countries: the 1995 Australia Study on healthcare quality had been drawn on the Harvard blueprint (Weingart et al. 2000), and a similar approach has been followed in a 1998 study on New Zealand public hospitals (Davis et al. 2002), a 1999–2000 English study (Vincent et al. 2001), a 2004 Dutch study (Zegers et al. 2009), and a 2005 study on Spanish hospitals (Ministero de Sanidad Y Consumo 2006), just to mention a few. Overall they assess a level of incidence of adverse events between 2.7% and 16.6% for the public systems and between 3.7% and 17.7% for the American system (Weiler et al. 1993; Andrews 2005). Around half of them are judged to be preventable[1]. The idea underpinning risk/error management solutions is that bad things do not happen to bad people (Reason 2000); rather "it is weak systems that create the conditions for error" (Department of Health 2002). According to this managerial awareness, the Department of Health in the UK has promoted initiatives to implement the error report system between local and national level. In the aftermath of the 1999 Institute of Medicine Report *To Err is Human*: *Building a Safer Health System*, many countries felt the urgent need to introduce a "safety culture" within their health systems (Barach and Small 2000). Improve error disclosure, a better interaction among different branches of the same health system, and homogenization of some basic medical procedures are among the undertaken steps. Both insurance companies and medical associations have stirred up such urgency. Additionally, it has been frequently recalled the importance of looking to "high reliability organizations" (i.e., aviation, nuclear power plants) in order to import or shape risk/error management policies that had been widely tested, especially to deal with near misses. Leape and Berwick (2005) undertook a study on the changes in hospital practice to improve safety in the USA 5 years after the publication of *To Err is Human*. Although a national inquiry is still missing, a local level analysis has been characterized by a decrease of adverse drug-related events and infections. Similar investigations (*how* the policy enforcement has effectively reduced *what*) would be desirable in other countries. An accurate analysis of the administrative costs of such policies is missing both at local and national level.).

However, the adopted definition of injury/adverse event, starting with the US studies, has been quite criticized. Identifying an injury as "any (negative) deviation from the expected outcome" and adverse event as "an unintended injury that was caused by medical management and that result in measurable disability" (Danzon 2000, p. 1352) has been judged either a too broad or a too narrow approach. Some authors, generally not economists, think that such definitions do not allow the inclusion of errors that are not associated to any injury either because the doctors were extremely lucky (and the patients too) or because the doctors could catch the mistake in time (again, a matter of luck) (Weingart et al. 2000). According to this interpretation, the findings should be regarded as the lower bound of a much striking number. Other authors, generally economists, disagree with the previous view, arguing against a loose definition used in the recalled studies and raising the problem of the efficient error, that for which the cost of precaution is equal to the expected damages. In other words, since we have not decided ex ante what is the efficient slot of errors related to the practice of a high risk activity as healthcare, we could not state for sure whether those numbers represent efficient or inefficient errors. Consequently, the empirical findings could be viewed as an exaggerated upper bound of a much less sensational phenomenon. Is the "preventable" number of adverse events really representative of inefficient errors? Does "preventable" mean efficiently preventable (precaution costs < potential benefits)? These questions still need to be addressed in practice.

Despite the fact that the two views are clashing, they are extremely useful to get a flavor of the range of factors we should evaluate and weigh when we try to empirically assess the incidence of iatrogenic injuries. A further contribution in this respect comes from a strand of less sound econometric analysis, run sometimes by physicians, which proposes an issue raised by other studies in the field: why injured patients do not always file a claim? Michael Rowe (2004), for example, supports with several case studies the evidence that the probability that a patient will file a suit will decrease whenever he/she has been told about the error. Paradoxically, wards – who are more exposed to the eventuality of error like Emergency Rooms, but where the doctors have a closer and constant supervision of the

[1] It might be worthwhile to mention the "risk management" approach on this point. Indeed a central issue is "whether negligent injuries are caused largely by occasional inadvertent lapses of many, normally competent providers or by a minority of incompetent, physicians and low quality hospitals" (Danzon 2000)

patient – tend to get less malpractice suits than other safer wards.

A final concern related to the assessment of medical errors is how we should judge the relationship between errors in medicine and quality of the healthcare service. This is in a way a sort of paradox. The actual improvement in the medical technology and a better knowledge of tackling endemic pathologies, considered fatal just a couple of decades ago, have transformed medical risk in two ways. On the one hand the new technology creates risks that did not exist previously (This belongs to the well-known path of human progress. The automobile invention had caused both development and accidents!), increasing also the skills required for its use. On the other hand the medical science progress makes the "time" factor crucially relevant to diagnose lethal pathology and provides solutions at the early stages of its development (i.e., cancer; Grady 1992). In other words, ceteris paribus, an increase in the service quality, at least in terms of adoption of technology, can hold an increase in risk of being suited and in expansion of the object of compensatory claims (For instance, a 2004 decision of the Italian Supreme Court established the right of compensation for iatrogenic *loss of chances* as a juridical independent category. Hence it is possible to file suit both for iatrogenic injury and iatrogenic loss of chances). On this respect a crucial role can be played by liability rules, which can or cannot favor the decision of adopting new technology.

The second strand of empirical literature is not always unambiguous as well and mainly concerns the efficacy of the policies adopted to decrease MM pressure. Decrease the pressure in this context means affecting the probability to be the target of a claim (the physician/hospital), the probability to compensate a claim (insurers), and the probability to get compensation (patient). The main target of these policies is to relieve healthcare practitioners and insurance companies from MM pressure so to decrease problems such as defensive medicine or the availability and affordability of insurance coverage. These goals should be pursued while the quality of the healthcare system is preserved or, at best, enhanced. The policies at stake are twofold: (1) reforms in the realm of tort law, not designed specifically to tackle MM problems, but that can directly or indirectly affect the behavior of the parties struggling with MM problems, and (2) reforms affecting the organization of the MM insurance market.

A first group of policies/reforms related to tort law has been the most studied in the USA from an empirical perspective. These policies have been distinguished in *direct* (i.e., caps on damage compensations) and *indirect* (i.e., pretrial screening panels) (Kachalia and Mello 2011; Kessler 2011). Although the evidence on the final impact of these policies on both liability and care-related measures is mixed, direct reforms seem to be associated to a stronger effect and in particular caps on damages are regarded among the most effective adoptable measures to decrease MM pressure[2].

Besides these two groups, tort reforms might include also the shift from negligence to no-fault liability and the adoption of form of enterprise liability. These reforms, differently from the previous, have often been adopted with specific reference to MM, as in the case of the UK (i.e., enterprise liability of hospitals; Fenn et al. 2004), Virginia, and Florida (i.e., no-fault system for severe birth-related neurological damages), as well as the Scandinavian countries (i.e., no-fault liability system for every medical injury). In the European cases the change in the liability regime was accompanied by a change in the MM insurance structure. It coincided with a change from a private/commercial to a public-taxpayers'-paid insurance coverage. The two shifts not always coincided. For example, Denmark adopted left the fault liability system for a no-fault system in 1992, but only from 2004

[2] However, the empirical results are ambiguous, and they depend (1) on the caps' target, as punitive damages, rather than economic or noneconomic damages, and (2) on the period of caps' introduction, the reforms implemented in the 1970s, in the 1980s, or in the 1990s. For a review see Kachalia and Mello (2011). Many European countries adopt schedules of noneconomic damages rather than caps, as in the case addressed by Bertoli and Grembi (2013)

the insurance system became completely public (also for private providers until 2013). So far, it has been difficult to empirically disentangle the effect of the change in the liability regime from the change in the insurance regime, so the attention has been focused especially on some elements of the latter, as in the English case (Fenn et al. 2007). The problems related to systems of public insurance as the case of the English *Clinical Negligence Scheme for Trusts* under a fault system or, for instance, of the Swedish *Patient Compensation Insurance* under a no-fault system are that public insurance can trigger common pool of risks more than private insurance. This means that often other institutional elements are requested to play a crucial role when adopting a public coverage, as for instance a proper monitoring mechanism on the flow of MM claims by a proper authority (Towse and Danzon 1999; Amaral-Garcia and Grembi (2014)).

A sound evaluation of the effects of a shift from negligence to strict liability has not been produced yet, and therefore, the main references are often translated by the experience of car accidents. Strict liability would allow for more liquidated claims, an average level of compensation lower than under alternative regimes, and a higher incidence of fatal accidents. However, given the peculiarities of the healthcare context, there are no strong arguments to infer that we should always expect the same results.

Changes at the fault regimes of MM remain one of the most debated issues lately. For instance, a recent document of the English Department of Health released on February 2014 (Department of Health UK 2014) addresses the concerns that the liability system did not foster medical innovation. Empirical analyses, which can help to address this and the use of additional tools to incentive efficient levels of precaution given specific institutional settings, are fundamental. Gathering and critically analyzing data on medical error remains priority one for many administrations (As underlined in the General Accounting Office (GAO) (USA) reports (GAO-04-128 T). The GAO has addressed several times the Congress to take initiatives in order to collect reliable data from the States regarding malpractice trends and effects).

## Cross-References

▶ Medical Liability
▶ Strict Liability Versus Negligence

## References

Amaral-Garcia S, Grembi V (2014) Curb your premium: the impact of monitoring malpractice claims. Health Policy 144(2):139–146

Andrews L (2005) Studying medical error in situ: implications for malpractice law and policy. De Paul Law Rev 54:357–392

Arlen J (2013) Economic analysis of medical malpractice liability and its reform. New York University public law and legal theory working papers. Paper 398

Barach P, Small SD (2000) Reporting and preventing medical mishaps: lessons from non-medical near miss reporting systems. Br Med J 320:759–763

Bertoli P, Grembi V (2013) Courts, scheduled damages, and medical malpractice insurance. Baffi center working paper. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2367218

Ministero de Sanidad Y Consumo (2006) National study on hospitalisation-related adverse events. ENEAS 2005. Report. Madrid

Danzon P (2000) Liability for medical malpractice, Chap 26. In: Newhouse J, Culyer A (eds) Handbook of health economics. Elsevier, New York, pp 1339–1404

Davis P, Lay_Yee R, Briant R, Ali W, Scott A, Schug S (2002) Adverse events in New Zealand public hospitals I: occurrence and impact. N Z Med J 115(1167):1–9

Department of Health (UK) (2014) Legislation to encourage medical innovation: a consultation. Available at https://www.gov.uk/government/organisations/department-of-health. Accessed Feb 2014

European Commission, D.G. Sanco (EC) (2006) Special eurobarometer medical errors

Fenn P, Gray A, Rickman N (2004) The economics of clinical negligence reform in England. Econ J 114:F272–F292

Fenn P, Gray A, Rickman N (2007) Liability, insurance, and medical practice. J Health Econ 26:1057–1070

GAO-04-128T (2003) Medical malpractice insurance. Multiple factors have contributed to premium rate increases, testimony before the subcommittee on wellness and human rights, committee on government reform, house of representatives. U.S. Government Printing Office, Washington, DC

Grady MF (1992) Better medicine causes more lawsuits, and new administrative courts will not solve the problem. Northwest Univ Law Rev 86:1068–1093

Hospitals of the European Union (HOPE) (2004) Insurance and malpractice, final report. Brussels, www.hope.be

Kachalia A, Mello MM (2011) New directions in medical liability reform. N Engl J Med 364:1564–1572

Kessler DP (2011) Evaluating the medical malpractice systems and options for reform. J Econ Perspect 25(2):93–110

Leape LL, Berwick DM (2005) Five years after *To Err is Human*. What have we learned? JAMA 293:2384–2390

Mello MM, Studdert DM, Brennan TA (2003) The new medical malpractice crisis. N Engl J Med 348:2281–2284

Nys H (2009) The factual situation of medical liability in the member states of the council of Europe, reports from the rapporteurs, conference 'The ever-growing challenge of medical liability: national and European responses', Strasbourg, 2–3 June 2008, pp 17–28

OECD (2006) Medical malpractice, insurance and coverage options, policy issues in insurance, no. 11

Reason J (2000) Human error: models and management. Br Med J 320:768–770

Rowe M (2004) Doctor's responses to medical errors. Crit Rev Oncol Hematol 52:147–163

Sage WM (2003) Understanding the first malpractice crisis of the 21st century. In: Gosfield AG (ed) Health law handbook. West Group, St. Paul, pp 549–608

Shavell S (1987) Economic analysis of accident law. Harvard University Press, Cambridge

Towse A, Danzon PM (1999) Medical negligence and the NHS: an economic analysis. Health Econ 8:93–101

Vincent C, Neale G, Woloshynowych M (2001) Adverse events in British hospitals. Preliminary retrospective record review. Br Med J 322:517–519

Weiler PC, Hiatt HH, Newhouse JP, Johnson WG, Brennan TA, Leape LL (1993) A measure of malpractice: medical injury, malpractice litigation and patient compensation. Harvard University Press, Cambridge

Weingart SN, McL Wilson R, Gibberd RW, Harrison B (2000) Epidemiology of medical error. Br Med J 320:774–777

Zegers MD, Bruijne MC, Wagner C, Hoonhout LHF, Waaijman R, Smits M, Hout FAG, Zwaan L, Christiaans-Dingelhoff I, Timmermans DRM, Groenewegen PP, Van Der Wal G (2009) Adverse events and potentially preventable deaths in Dutch hospitals: results of a retrospective patient record review study. Qual Saf Health Care 18:297–302

# Mercantilism

Günther Chaloupek
Austrian Chamber of Labour, Vienna, Austria

## Definition

**Mercantilism** is a system of economic policy and a corpus of economic doctrines which developed side by side from the sixteenth to the eighteenth century. The main goal was to increase a nation's wealth and power by imposing government regulation to promote the nation's commercial interests by maximizing exports and limiting imports.

## Mercantilism, Colbertism, Cameralism

**Mercantilism** is a system of economic policy and a corpus of economic doctrines which developed side by side from the sixteenth to the eighteenth century. As a theory, mercantilism marks the decisive step in the emancipation of thinking about economic phenomena from scholastic theology to political economy and economics as a social science of its own. With respect to economic policy, mercantilism took a variety of different forms according to the different political, economic, and social conditions prevailing in European states during the early modern period. As a consequence, there are national variants of mercantilist economic literature focusing on different aspects of the economic process.

In a more general perspective of history of ideas, mercantilism as a body of theoretical doctrines as well as a system of economic policy is part of the emergence of a rationalist worldview with its understanding of natural phenomena in terms of cause and effect, instead of purpose inherent in the substance of things. Central to the new worldview is the concept of law as a force independent of human intention applicable to external physical nature and to human nature. In the spirit of the Baconian sentence *scientia est potentia*, knowledge of such laws brings with it the power to influence the course of events according to desired goals.

The early modern period was the time when nation states took shape on the European continent and in England. It was in this context in which the new political and economic doctrines acquired practical relevance. As a consequence, the focus of mercantilist authors was primarily on relations between states and on collectives within states (Pribram 1983, p. 83). That the perspective of the individual agent was only relevant in this context explains the contempt of classical and

Günther Chaloupek has retired.

neoclassical economics for mercantilism, whereas other schools of thought, such as the German Historical School or Keynesianism, are more prepared to acknowledge the merits of some of its doctrines.

None of the mercantilist authors has produced a compact theoretical system of the working of the economy. Rather, "the economics of mercantilism" is an ex post construct of history of economic theory which has identified a number of core issues and policy problem upon which the debates centered. Among those issues, the balance of trade, or, in a wider sense, the balance of payments of a nation is the most important one. But it would be wrong to identify the emphasis on achieving an export surplus with the opinion that national wealth amounts to nothing else than the accumulation of treasure, as was and still is often suggested. Rather, an improvement of the balance of payments was in various ways seen as a policy strategy to develop a country's economic potential and thereby enhance its power in the context of rivalry among European nations. The relevance of specific ramifications of the external balance issue, such as import restrictions, export promotions, creation of legal monopolies, acquisition of colonies, exchange controls, varied according to the different circumstances under which individual nations sought to establish themselves in the competition of European powers. The key role for economic development which is assigned to trade extends beyond external to domestic trade and the means of its promotion. Increasing awareness of interdependence of a multitude of policy instruments lead to the formulation of policy concepts with the claim to regulate the economy as a whole (Colbertism). Given the prevalence of policy aspects in mercantilist literature, for some historians of economic theory the main relevance of mercantilism consisted in providing a political doctrine for formation of national states through the replacement of the local and regional government by the central government (Schmoller 1883) or as a system of power politics (Heckscher 1935). In contrast, Schumpeter (1954) carefully elaborated the substantial contributions of mercantilist authors to theoretical economic analysis.

If mercantilist economics had an early start in England, this can be attributed to the fact that the government had reached a comparatively high degree of centralization at the beginning of the Modern Age. In addition, consolidation of the central government coincided with the buildup of a colonial empire with its rapidly expanding trade. Conflicting interests of commercial capital are reflected in books and pamphlets whose authors often "clothed their views in the garb of a policy designed to strengthen the nation" (Roll 1942, p. 58f). Gerard de Malynes (ca. 1555–1643) warned against the loss of precious metal due to the fall of the exchange rate below silver parity which he attributed to the lack of foreign exchange controls and to the privilege of the East India Company for limited export of bullion. This focus on exchange rate was contradicted by Edward Misselden (ca. 1608–1654) who introduced the concept of balance of payments which should be seen as true indicator whether trade was beneficial for a country. A surplus should be achieved through promotion of exports and discouragement of imports, especially imports of luxuries. Interests of commercial capital found their "fullest expression" (Roll, p. 75) in the work of Thomas Mun (1571–1641). Building upon Misselden's balance of payments theory, Mun argued that the export surplus augmented the capital ("stock") that could be invested in trade and production and thus enhanced wealth and power of the nation. The debate between Josiah Childs (1630–1699) and Sir Dudley North (1641–1691) focused on the interest rate as a possible cause for the depression which hit England during the period of naval warfare against the Netherlands. Childs argued that the high interest rates which English merchants had to pay in comparison with Dutch competitors were responsible for depression and called for limits to be enforced by the state. North reversed the argument by saying that it was an increase in the volume of trade which would lead to an increase of the quantity of money and thus lower the rate of interest. If North proposed to do away with measures of trade protection and prohibition for that purpose, this foreshadows the end of mercantilist thinking in England. Of the important contributions of Sir William Petty (1623–1687), mention should be made of his *Political Arithmetick* in which he advocated the use of *number*, *weight*, and *measure* in debates about economic issues – an early example of quantitative empiricist method in the attempt to establish laws of nature. English mercantilist authors extensively reflected also on prices and wages, taxation, population, etc., in the context of their respective causes. First attempts to assemble these elements of economic analysis into a coherent system from a mercantilist perspective, i.e., from the perspective of the state, were undertaken by Richard Cantillon in his *Essai sur la nature du commerce en general* (1755, originally written in English) and by Sir James Steuart in his *Principles of Political Economy* (1767).

If the **Netherlands** are often cited as model by English mercantilists, this is due the close identification of the country's government with the interest of the merchant class. While the political influence of the landed aristocracy in politics was still strong in England, in the seventeenth century Holland is the merchant state *par excellence*. The interests of Dutch merchants were best served by free trade for which the legal sciences (Hugo Grotius) provided the best arguments.

Following a path different from England in its formation of a national state, in **France** political and administrative power had been concentrated the hands of the king ruling as absolute monarch. French mercantilism was above all a comprehensive system of administration which sought to develop the economic powers of the country through "retablissement des manufactures," advocated by Bartéhelemy Laffemas (1545–ca. 1612) who served as *controleur* under King Henri IV (Sommer 1920/1925, p. 29). A similar approach was pursued by Antoine de Monchrétien in his *Traicté de l'oeconomie politique* (1615) in which this term appears for the first time. French mercantilism was fully developed in practice, much less in theory, during the reign of Louis XIV by his finance minister Jean Baptiste Colbert (1619–1683). Under Colbert, population policy was adjusted to the aims of power policy, external trade was conducted as a kind of warfare against England and the Netherlands, the acquisition of precious metal was proclaimed as the main objective of external trade, all measures of commercial policy were ruled by the endeavor to promote exports and prevent imports of final products (Pribram 1983, p. 51). "Colbertism" came to be used as synonym for mercantilism. It was the model for the "cameralist" system established in Austria in the eighteenth century.

In the territorial states of the **German Empire**, mercantilist practice and theory appears in different forms. In Prussia and in the Austrian monarchy mercantilist policies were deliberately used to create a unified internal market, thereby also strengthening political control of the central government over the heterogeneous provinces. Among the cameralist authors who offered their advice to the Habsburg Emperors, Johann Joachim Becher (1635–1683) made the most important theoretical contributions with his doctrine of market forms which distinguishes between monopoly, "polypolium," i.e., free competition with free access to markets, and "propolium," by which he means various kinds of restrictive or speculative practices. Rejecting all three forms, Becher pleads for some kind of organized and supervised competition. Becher assigns a key role to commerce in the efforts to make the economy more dynamic and recommends the foundation of sectoral trading companies by the state as instrument to encourage industrial activities. Philipp Wilhelm von Hörnigk, in his book *Österreich über alles, wenn es nur will* (1684) drew up a comprehensive program to create a national economy in the Habsburg crownlands. Hörnigk's tract as well as later cameralist literature highlight the essentially defensive orientation of Austrian mercantilist policies, which aim at an improvement of the external balance through a strategy of import substitution, in contrast to offensive export promotion by Western European states. Veit Ludwig von Seckendorff (1626–1692) who served as chancellor in the small state of Sachsen-Weimar puts significantly more emphasis on general conditions of production, such as reliability of legal framework, a stable monetary system, moderate taxation, education and training, investment in infrastructure, improvement of sanitary conditions, etc., while on the other hand he devotes much less attention than Becher to interventionist measures of promotion of trades.

Johann Heinrich Gottlob Justi (1717–1768), the most important cameralist author of the eighteenth century, came close to recommending autarky. If external commerce was beneficial, it was not a necessity. "An empire may be very powerful, wealthy and flourishing without having external commerce with other peoples; alone, never can there be a state of such a character if its manufactures and industries are not flourishing." It is in this context where Justi developed his central concept of *Universalkommerz* for which Colbert's system served as model: "The sovereign has to direct all trades according to the needs of the country and to the requirements of its external commerce, of the promotion and augmentation of the livelihood of its subjects, and – in brief – of the general welfare." Earlier than in England, chairs were established at German universities for mercantilist economics under the title "cameral sciences" or "police sciences," the first one 1727, in Halle, Prussia. German textbooks, such as Justi's *Grundsätze der Polizeywissenschaft* (2 vols., 1756) and Joseph von Sonnenfels' *Grundsätze der Polizey, Handlung und Finanz* (3 vols., 1765ff), are handbooks for practical policy, rather than syntheses of analytical knowledge.

Well into the twentieth century, popular perceptions of mercantilism have been shaped by Adam Smith's attack on what he called the "commercial" or "mercantile" system for its comprehensive regulations of imports and exports which he considered undue restrictions of freedom and obstacles to augment the wealth of a nation. Meanwhile, economic history and history of economic theory have corrected Smith's verdicts in important respects and produced a more balanced picture of the merits and errors of this early stage of economics. It is widely recognized that free trade cannot under any circumstances be considered the best strategy to foster economic development, which can be supported by well-designed, temporary state interventions. On the other hand, a strategy of export-led growth which can turn into some kind of "new mercantilism" has remained a powerful temptation not only for newly industrializing nations, with the risk of accumulating large-scale international imbalances. It appears that important issues of mercantilist economics will remain relevant in the twenty-first century.

## References

Heckscher EF (1935) Mercantilism. Allen & Unwin, London

Pribram K (1983) A history of economic reasoning. Johns Hopkins University Press, Baltimore/London

Roll E (1942) A history of economic thought. Prentice-Hall, New York

Schmoller G (1883) Das Merkantilsystem in seiner historischen Bedeutung. In: Jahrbuch für Gesetzgebung, Verwaltung und Volkswirtschaft, vol VIII

Schumpeter JA (1954) History of economic analysis. Allen & Unwin, London

Sommer L (1920/1925) Die österreichischen Kameralisten in dogmengeschichtlicher Darstellung, Heft XI und XII der Studien zur Sozial-, Wirtschafts- und Verwaltungsgeschichte, Carl Grünberg C (ed), Vienna, reprint Scientia Verlag, Aalen 1967

# Merchants' Law

► Lex Mercatoria

# Merger Control

Tim Reuter
RBB Economics, Brussels, Belgium

**Abstract**

Merger control is at the heart of competition law institutions throughout the world. Firms, before they complete a merger or an acquisition, are required to get the transaction approved by competition authorities. The objectives of merger control, limiting the accrual of market power and protecting the welfare-generating competition forces of the market, are well in line with economic theory. Economic theory has identified several types of effects that can arise from a merger, in particular unilateral and coordinated horizontal effects that result from an elimination of competition between firms supplying substitutable goods, and non-horizontal and conglomerate effects that result if firms are active on vertically or otherwise linked markets. We discuss the reasoning behind these effects and how they are assessed by competition authorities. We also discuss market definition as a first step in the assessment of mergers and the legal framework under which mergers are controlled.

## Introduction: Objective of Merger Control and a Classification of Merger Effects

In most jurisdictions throughout the world, merger control procedures are a central element of competition law institutions. The objective of merger control is to limit the accrual of market power and to maintain the process of competition in the market in order to protect (consumer) welfare. To achieve these ends, merger control procedures must anticipate the competitive effects mergers will bring and establish enforceable rules according to which mergers are blocked or approved. In certain jurisdictions, merger control rules might have objectives other than maintaining competition, for example public interest objectives. Such objectives are however usually not considered as part of competition law in the narrower sense. We will not discuss these aspects in this note.

From an economic theory perspective, the first general theorem of welfare economics states that Pareto efficient allocations will only be achieved if firms behave as price takers, i.e., act as if their own output decisions have no influence on price or, in other words, have no market power. In comparison to such models with atomistic firms, models with imperfect competition feature losses of consumer welfare. The objectives of merger control are hence in line with a welfare maximization objective. It is also general consensus that it is preferable to limit the accrual of market power via merger control over regulating firms not to exercise their market power once acquired (though competition laws have provisions to also limit the exercise of market power).

While an increase in market power leads to a lessening of (consumer) welfare from an economic theory viewpoint, mergers may be conducted without any effect in market power. They may even entail procompetitive effects. For example, it may be easier to generate economies of scale or scope after a merger. Mergers may also be conducted to increase productive efficiency or to merge two complementary businesses. The trade-off between allowing firms to realize efficiencies and allowing them to accrue market power has been denominated as “Williamson trade-off” after a famous paper by Williamson 1968. As such, any merger control regime that would block mergers per se would be overly interventionist. Instead, in competition law regimes throughout the world, case-by-case assessments are conducted, in order to predict the competitive effects the mergers entail.

From an economics perspective, mergers can lead to consumer harm in several ways. First, harm can arise from a lessening of competition if firms merge that are active on the same market, i.e., supply substitutable products. Such harm is labelled a horizontal effect. Two types of horizontal effects are distinguished: The first type arises if the merged entity has an incentive to increase prices as a result of a reduction of direct competitive pressure by the removal of a competitor. This is called a “unilateral effect.” The second type arises if the merger facilitates for the remaining firms in the market to coordinate on some anticompetitive behavior. This is called a “coordinated effect.”

A second category of potential harm is called vertical effects, i.e., effects that arise if the firms are active on related segments of the same supply chain. Vertical effects typically require that some competitor is foreclosed from accessing either inputs or customers. Finally, conglomerate effects can arise. Under conglomerate effects, merging firms neither supply competing goods nor are they vertically linked. Harm can arise, because the merged firm can leverage its broader scope (for instance by bundling complementary goods). As a rule of thumb, horizontal effects are the most frequent concern. Vertical and conglomerate mergers are less likely to entail anticompetitive effects and even often give rise to procompetitive effects.

## The Legal Framework of Merger Control

Before discussing the substantive economic assessment of mergers, based on which it will be decided whether a proposed merger is prohibited or not, we describe in this section certain procedural aspects of merger control. We focus on aspects that are common to most jurisdictions throughout the world.

First of all, usually not all mergers are assessed by competition authorities. Depending on the jurisdiction, usually some notification thresholds exist, below which mergers do not need to be cleared. These can be traction volume based (for example, in the USA, generally transactions in excess of 78.2 million USD are notifiable) or turnover based (under European Union law, mergers are notifiable if the combined worldwide turnover exceeds 5,000 million EUR or if the EU turnover of at least one concerned entity exceeds 100 million EUR). In most countries, notification means that a transaction must be approved before it is completed (where gun-jumping fines may be applied).

Regarding the scope of activities undertaken, the word merger control might suggest a too narrow definition of what it covers. In reality, in some jurisdictions, also concentrations short of a merger or an acquisition fall under merger control rules. For example, in the European Union, joint ventures have to be notified under certain conditions, and in some jurisdictions, also minority shareholding acquisitions have to be notified.

The role of authorities and courts can also differ fundamentally between jurisdictions. For example, under the European Union Merger Regulation, the European Commission can challenge mergers directly. Courts play only a role should any party (including third party complainants) appeal the European Commission's decision. In the USA in contrast, mergers are investigated by the Department of Justice or the Federal Trade Commission. Should they decide to challenge a merger however, they have to seek an injunction in front of a court. Either way, usually precise time tables govern the merger control procedures.

Finally, the choice of competition authorities (or courts where applicable) is not limited to a choice between approving a merger and blocking it. Firms can also propose to remedy the concerns raised by competition authorities. These can be structural, i.e., the merging parties agree to divest certain activities (where the divestment transaction may be itself subject to merger control) or behavioral, i.e., the merging parties commit themselves to some postmerger behavior as agreed with the competition authorities.

## The Substantive Assessments of Mergers

As discussed in the Introduction, the objective of merger control comes from the notion that restricting the accrual of market power enhances welfare. To implement this into enforceable rules, different legal test have been established according to which mergers are can be blocked. While traditionally, in many jurisdictions, the legal test for prohibiting a merger was whether it would establish or strengthen a dominant position; nowadays most jurisdictions have adopted a more nuanced test.

For example, according to the European Union Merger Regulation enforced by the European Commission (and most member states have similar rules), a merger is to be prohibited if it causes a *significant impediment of effective competition* (the so-called SIEC test), in particular if it gives rise to a dominant position (until 2004, the creation or strengthening of a dominant position was the sole reason for a prohibition under European merger control). Similar, in the USA and the UK, a merger is to be prohibited if it gives rise to a *substantial lessening of competition* (the so-called SLC test). While the SIEC test explicitly mentions the dominance criterion as a particular example under which mergers can be blocked, the SIEC and SLC tests are generally considered to be similar and remaining differences for practical purposes are nuanced.

In the following, we describe how competition authorities economically assess and decide whether a proposed merger is to be cleared or blocked (and whether remedies might be appropriate) according to the SIEC or SLC test. This description is based on guidelines that various competition authorities have published to provide guidance on their assessment of mergers. In particular, the European Commission published a "Commission notice on the definition of the Relevant Market for the purposes of Community competition law" (1997), "Guidelines on the assessment of horizontal mergers" (2004), and "Guidelines on the assessment of non-horizontal mergers" (2008), while the US Department of Justice published "Horizontal Merger Guidelines" (2010) and "Non-horizontal merger guidelines" (1997).

A first step in the assessment of likely effects is the definition of the relevant market. Then, if necessary, horizontal and nonhorizontal effects are assessed. We discuss each of these steps in turn.

## Market Definition

The first step of the assessment of mergers under both the SIEC and the SLC test is the definition of the relevant market. The objective of defining the relevant market is to identify the competitive constraints firms face and to provide a framework for the competitive assessment. In particular, defining the market allows the measurement of market shares and other concentration-based statistics.

The definition of a relevant market comprises the combination of both a product market definition and a geographic market definition. With respect to the product market definition, a given product market is constituted by all products that are interchangeable or substitutable by consumers. The substitutability is to consider all product characteristics, prices, and the intended use of the products. With respect to the geographic market definition, the market comprises all areas in which the conditions for supply and demand for the given services are homogenous and can be distinguished from other areas in which conditions of competition are sufficiently different.

In practice, to assess the relevant market definition, the Hypothetical Monopolies Test ("HMT") is implemented. The HMT is a sequential test that starts with the narrowest possible candidate market and asks whether it would be profitable for a hypothetical monopolist in that market to implement a small but significant nontransitory increase in price ("SSNIP"). If so, it is worth monopolizing the market, i.e., a monopolist in this market does not face significant competition from outside this market and the relevant market is found. If it is not profitable to increase prices, the hypothetical monopolist faces some competition from outside the market. The candidate market definition is discarded. As next sequential step then, the candidate market definition is widened and the SSNIP question is reapplied assuming a hypothetical monopolist for the wider market. The process is reiterated until a market is wide enough such that the price increase is profitable. For this market, the hypothetical monopolist faces no significant competition from outside the market.

The HMT can be implemented using several empirical techniques to answer the question whether a price increase is profitable for the hypothetical monopolist, for example critical loss analysis, demand estimation, analysis of geographic sales patterns, analysis of price levels and price correlation, and stationarity analysis (see Gore et al. 2013 for a description of these techniques).

Finally, it should be noted that certain commentators have argued that in some cases, market definition may not be a necessary step and that economic effects can be directly assessed (see Kaplow 2010). This notion seems to be better-received in some jurisdictions (in particular in the USA and the UK) than in others.

## Horizontal Effects

Horizontal effects may arise under mergers of firms that supply substitutable goods, i.e., firms that are competing directly with one another. Two types of horizontal effects are distinguished. First, a merger may eliminate direct competition between the merging firms (called a "unilateral effect" or a "noncoordinated effect"). Second, a merger may lessen or impede competition by making coordinated behavior more likely, i.e., behavior of multiple firms that is profitable for each of them only as a result of the reactions of the other firms, i.e., (tacit or explicit) collusion (called a "coordinated effect"). We discuss unilateral and coordinated effects in turn.

### Unilateral Effects

A first step in the assessment of unilateral effects is usually the calculation of market shares and other concentration measures, such as the Herfindahl concentration index ("HHI"). These measures are in particular useful evidence if products are homogenous, if capacity can be easily expanded, and if goods are traded in a spot-market fashion. In these cases, concentration measures often constitute a presumption for the competitive effects of the merger (for instance, the European Commission presumes that there is a significant impediment of effective competition if the combined market shares of the merging parties are above 40%; in US enforcement, mergers involving an HHI increase of more than 200 points are presumed as likely to enhance market power). In some cases, competition authorities conduct price-concentration analyses, investigating whether there is an empirical relationship between the concentration in a market and its given price level.

However, concentration-based measures are in many cases only the starting point of the competitive assessment of mergers. Further elements are particularly relevant if products within the market are not homogenous, if firms in the market face capacity constraints, or if customers procure the goods in a nonstandard way, for example, via auctions.

Concentration-based criteria are unable to inform competition authorities in a precise way about the competitive constraints firms face, in particular in markets, in which products are differentiated. For instance, if the market definition has revealed that the three products A, B, and C are in the relevant market, and if B and C have the same market shares, the concentration-based criteria presume that B and C exert the same competitive constraint on the pricing of product A. However, consumers may actually be more willing to substitute product A with product B rather than with product C. In such a case, competition authorities investigate typically the precise substitutability between products to determine the relative importance of all products in the defined market as competitive constraints to the products of the merging firms.

For instance, the European Commission investigates the closeness of competition of goods within the relevant market. One piece of evidence for this purpose can be an analysis diversion rates, e.g., the fraction of unit sales lost by a price increase of a certain goods that would be diverted to the sales of a second product. A high diversion ratio from a product of one the merging firms to a product of the other merging firm is an indication of a high likelihood of an anticompetitive effect (ceteris paribus the market shares).

Competition authorities use a number of empirical techniques to assess the competitive constraints imposed by particular products within the relevant market, for example, diversion rates can be assessed using survey evidence, customer switching data from firms' market intelligence databases (e.g., mobile number portability data in the case of mobile network operator mergers – see European Commission 2016), assessments of natural experiments (e.g., switching as the result of a closure of a plant – see Coate 2012), and analysis of win/loss data (see Botteman 2006). Advanced tools to estimate the constraints of particular products are upward pricing pressure tests (see Farrell and Shapiro 2010) and merger simulation models (see Budzinski and Ruhmer 2009).

Other factors that play a role in the assessment of unilateral effects are entry, capacity constraints of suppliers, countervailing or buyer power, efficiencies generated by the merger, product repositioning, failing firm defense, effects of the merger on innovation and product variety, and a loss of potential competition (e.g., one merging party planning to enter a given market where the other is already active). These factors may be treated differently in different jurisdictions however. For example, Brouwer (2008) points out that the main difference between EU and USA merger policy lies in the greater scope for efficiency arguments in the USA.

### Coordinated Effects

The notion of coordinated effects relies on the idea that a merger can make (tacit) collusion more likely to arise or can increase its effectiveness. As is well established by economic theory, collusive situations are difficult to establish for firms, because in a static setting, firms have generally an incentive to deviate from such situation by competing more aggressively. However, in a dynamic setting, collusive situations can be stable such that no firm has an incentive to deviate.

For the implementation in merger control, coordinated effects are assessed by judging whether the relevant market is generally vulnerable to coordinated conduct and whether the vulnerability increases by the merger.

For the vulnerability of the market in general, it is usually assessed (i) whether the firms in the market are likely to reach a common understanding of the terms of the coordination, (ii) whether firms can monitor whether all firms adhere to the coordination conduct, (iii) whether credible deterrent mechanisms are available that prevent firms from deviating from the coordination, and (iv) whether outside firms (competitors or customers) have ways of destabilizing the coordination.

As shown in economic theory, (tacit) collusion in dynamic settings often has many equilibria (see Friedman 1971). It is less clear which equilibrium may be a focal point and how firms coordinate on a specific tacit collusion equilibrium to play. This is reflected in the assessment of coordinative effects that firms must reach an understanding of the terms of the coordination conduct. While in some markets, for example, a market with few competitors, where competitors are symmetric, with homogenous products and that is not strongly affected by entry and innovation, reaching such an understanding may be simple. In markets which are more complex, reaching a tacit understanding is less likely. Hence, merger control takes these market characteristics into account when assessing coordinative effects.

The ability of firms to monitor and punish deviating firms depends on market characteristics such as transparency and stability of demand conditions (i.e., to determine whether an unexpected market outcome is result of deviating behavior or result of demand fluctuation), frequency and concentration of orders (it becomes more difficult to punish if orders are lumpy), and whether firms compete in a single- or multimarket setting (as punishment can also occur on markets other than the one on which coordination takes place).

Finally, competition authorities take reactions by other firms into account, for example whether competing providers that do not take part in the coordination have the ability to increase capacity, whether third parties can enter in the market and whether customers have countervailing power that can destabilize the coordination.

For the facilitation of coordination by the merger, after all what matters in whether coordination becomes more likely or more effective by the merger, the reduction of the number of firms active can be enough if market concentration is sufficiently high. However, even if the merger does not cause a significant increase in concentration, coordinative effects can be found, for example, if one of the merging firms is found to be a 'maverick,' i.e., a firm that is known to be disruptive to the market, e.g., by pricing aggressively or by innovating regularly.

## Nonhorizontal Effects

A merger can also affect market outcomes if the merging firms do not compete on the same market, but are active on different levels on the same supply chain (i.e., a vertical merger) or are active on different markets that are somehow related, for example, two markets for complement goods (i.e., a conglomerate merger).

In general, nonhorizontal mergers tend to be less likely to raise competitive concerns than horizontal mergers, because unlike the latter, they do not lead to a loss of direct competition between the merging firms. Often vertical mergers solve inefficiencies in the supply chain and are hence procompetitive; for example, they reduce double-marginalization (i.e., lower downstream prices, because the merged entity internalizes increased upstream profits from an expansion of output if downstream prices decrease), decrease transaction cost, or reduce prices for complement goods.

However, vertical mergers may not only solve inefficiencies and lead to procompetitive effects, but can under certain circumstances have anticompetitive effects as well. Two types of such anticompetitive effects are input foreclosure and customer foreclosure. Input foreclose relies on the notion that a firm active on a downstream market, by vertically integrating upwards, can start supplying its competitors on the downstream market on unfavorable terms only, thereby weakening downstream competition and increasing downstream prices. Customer foreclosure, on the other hand, means that a firm on an upstream market, by integrating downward, can shrink the customer base of its upstream competitors, thereby decreasing the ability of the upstream competitors to compete. In result, upstream prices may increase, whereby competition on the downstream market is hurt, leading to an increase of downstream prices.

Whether it is for the merging firms possible and profitable to foreclose competitors from accessing inputs or customers is however dependent on the specific market characteristics. For example, a downstream competitor cannot be foreclosed by a vertical merger if the merged entity does not have sufficient market power in the upstream market, because the downstream competitors can get access to the input from competing providers (at competitive terms). Similar, while input foreclose may increase the integrated firm's profits on the downstream market, it pays a cost for the foreclosure by lost profits resulting from less sales on the upstream market. As such, the merged entity might not have an incentive to engage in foreclosure, for example, if upstream margins are high and downstream margins are low.

In a similar vein, customer foreclosure is not necessarily possible and profitable for vertically integrated firms. First of all, for customer foreclosure to be possible, the integrated firm must have a sufficient degree of market power in the downstream market, such that competing upstream providers may actually face a significant loss of sales opportunities. Second, the loss of customers for the upstream competitors must entail that they are not able to compete any more effectively, which requires some form of economies of scale or scope for the upstream providers or that they operate at or close to minimum efficient scale. The profitability of consumer foreclosure depends on a similar trade-off as that of input foreclosure: By customer foreclosure, the integrated firm can lower competition and raise prices on the downstream market. To do so, it has however to carry additional costs on the upstream market (by eventually not procuring from the cheapest provider).

Foreclosure concerns can also matter in conglomerate mergers, i.e., for mergers in which the merging firms are active neither on the same market, nor on vertically linked markets. Similarly to vertical cases, conglomerate mergers will often entail procompetitive effects, for example, when the merging firms sell complementing goods (because they will lower prices for both goods, taking into account that a price decrease for one good will trigger an increase in demand for the complement). As well similar to vertical mergers, they can however trigger foreclosure concerns. For example, by tying or bundling two goods together, the conglomerate firm may leverage market power it holds in one market to another market to achieve above-competitive prices in that market. However, the firm will only have the possibility to leverage the power of one market to the other, if its market power in the first market is sufficiently high and if a large enough share of the consumers who buy one good also want to buy the second. The profitability of such foreclosure depends on the trade-off between losses in the market with market power (because some customers might refrain from buying this good if it is bundled to the other good) and gains in the market to which the power is leveraged.

For all types of foreclosure, competition authorities are hence comprehensively assessing market power of the integrated firm in at least one market, the cost of leveraging the market power that is entailed on the same market, and the benefit the leveraging causes on the second market.

## Cross-References

- ▶ Abuse of Dominance
- ▶ Horizontal Effects
- ▶ Market Definition
- ▶ Merger Remedies
- ▶ Type-I and Type-II Errors

## References

Botteman Y (2006) Mergers, standard of proof and expert economic evidence. J Compet Law Econ 2(1):71–100

Brouwer M (2008) Horizontal mergers and efficiencies; theory and anti trust practice. Eur J Law Econ 26(1): 11–26

Budzinski O, Ruhmer I (2009) Merger simulation in competition policy: a survey. J Compet Law Econ 6(2): 277–319

Coate MB (2012) The use of natural experiments in merger analysis. J Antitrust Enforc 1(2):437–467

Department of Justice and Federal Trade Commission (1997) Non-horizontal merger guidelines, https://www.justice.gov/sites/default/files/atr/legacy/2006/05/18/2614.pdf

Department of Justice and Federal Trade Commission (2010) Horizontal merger guidelines, https://www.ftc.gov/sites/default/files/attachments/merger-review/100819hmg.pdf

European Commission (1997) Commission notice on the definition of the relevant market for the purposes of community competition law. Off J Eur Union C 372:155–163

European Commission (2004) Guidelines on the assessment of horizontal mergers under the council regulation on the control of concentrations between undertakings. Off J Eur Union C 31:5–18

European Commission (2008) Guidelines on the assessment of non-horizontal mergers under the council regulation on the control of concentrations between undertakings. Off J Eur Union C 265:6–25

European Commission (2016) Recent developments in telecoms mergers. Compet Merger Brief 3(2016): 1–8

Farrell J, Shapiro C (2010) Antitrust evaluation of horizontal mergers: an economic alternative to market definition. BE J Theor Econ Pol Perspect 10(1), Article 9

Friedman J (1971) A non-cooperative equilibrium for Supergames. Rev Econ Stud 38(1):1–12

Gore D, Lewis S, Lofaro A, Dethmers F (2013) The economic assessment of mergers under European competition law. Cambridge University Press, Cambridge

Kaplow L (2010) Why (Ever) define markets? Harv Law Rev 124:437–440

Williamson O (1968) Economics as an anti-trust defense: the welfare trade-offs. Am Econ Rev 58(1):18–36

# Merger Remedies

Patrice Bougette
Department of Economics, Université
Côte d'Azur, CNRS, GREDEG,
Nice, France

### Abstract

Merger remedies are used by competition agencies to prevent the harm to the competitive process that may result as a consequence of a merger. They allow for the approval of mergers that would otherwise have been prohibited, by removing the anticompetitive concerns that a given transaction may pose to competition. First, we present the typology of merger remedies generally used. Second, we analyze the link between the size of the offered remedies and the level of efficiency gains announced in a context of asymmetric information. Third, we summarize the results of several retrospective merger studies in which remedies have been used.

Competition agencies use merger remedies when a notified merger is likely to raise some competitive concerns. In this case, the merging firms may make a remedial offer, which may be accepted or rejected by the agency. In this entry, we derive examples mainly from the European Commission (hereafter "the EC") even though most mechanisms can be found in any competition agencies worldwide.

Actually, remedies are relatively less used with respect to the total number of notified merger proposals to agencies. According to the EC's data, less than 10% of notified mergers are eventually conditioned upon merger remedies (see the EC's website for detailed data on the European merger control, http://ec.europa.eu/competition/mergers/statistics.pdf). The majority are approved without any conditions. However, these remedies may be found in large or complex merger cases. In the absence of remedies, such mergers would be rejected.

## Types of Merger Remedies

Two types of remedies are commonly used by competition agencies, structural and behavioral remedies. We first detail them and then discuss the pros and cons of each in terms of implementation.

Structural remedies refer to a transfer of ownership rights. For instance, merger firms commit themselves to selling a portion of their assets to one or several buyers. The selected assets may be located in markets where the level of concentration is very high. These assets are often the merging parties' overlap. For instance, in 2016, the EC approved the acquisition of beer group SABMiller by Anheuser-Busch InBev provided that, among other commitments, the new group divested SAB's brands Peroni, Pilsner Urquell, and Grolsch (*ABI*/*SAB* merger case, M.7881; see the EC press release, "Mergers: Commission approves AB InBev's acquisition of SABMiller, subject to conditions," 24 May 2016).

By nature, structural remedies are difficult to reverse and may largely apply to horizontal practices. The sale of an independent activity operating in the market appears more effective than a transfer of heterogeneous assets from the two merging firms (see the European Commission (2005)'s study for a categorization of these remedies).

Behavioral remedies may be defined as constraints on the property rights of the new entity. These commitments do not affect the market structure. They include third-party access to infrastructure or technology, or the parties commit themselves to putting an end to any exclusive vertical agreements. For instance, in the context of the acquisition of Arianespace by Airbus Safran Launchers (ASL), a joint venture between Airbus and Safran, the EC had concerns that the transaction would give rise to flow of potentially sensitive information between the two companies. This would have been at the expense of rival satellite manufacturers and launch service providers. Among other behavioral remedies, the companies committed not to share information about third parties with each other (*ASL*/*Arianespace* merger case, M.7724; see the EC press release, "Mergers: Commission approves acquisition of Arianespace by ASL, subject to conditions, 20 July 2016).

In general, competition agencies have a preference for structural commitments. Indeed, a behavioral remedy involves direct monitoring costs, whereas in the case of a structural remedy, once transferred, the assets no longer need special monitoring. However, in this case, it may be useful to set up a follow-up mechanism until the assets are sold. For instance, with regard to complex merger cases, an independent trustee may be appointed to monitor the smooth transfer of assets and strict compliance with time limits (see the Commission Notice on remedies acceptable under the Council Regulation (EC) No 139/2004 and under Commission Regulation (EC) No 802/2004 Official Journal C 267, 22.10.2008, p. 1–27, http://ec.europa.eu/competition/mergers/legislation/files_remedies/remedies_notice_en.pdf).

Nevertheless, the preference for a structural option may be flexible and adapted to the case context. First, the two types of remedies are not necessarily exclusive, behavioral commitments may be useful to a structural solution (see, e.g., the EC *Orange*/*Jazztel* merger case, M.7421, 19 May 2015). Then, in some cases, perhaps no competitor is interested in the proposed assets, in which case the agency turns to a behavioral commitment. In addition, in highly changing market circumstances, nonstructural remedies may constitute a more flexible and a revisable solution (Motta et al. 2007). Lastly, transfer of assets may facilitate collusion if they lead to a more symmetric industry structure. Compte et al. (2002) show that due to the use of remedies, coordinated effects emerged in the context of the 1992 *Nestlé–Perrier* merger case.

## Efficiency Gains and Asymmetric Information

A number of theoretical models have focused on the effects of structural merger remedies. For instance, such remedies enlarge the scope for approvable mergers in the presence of merger synergies. Nonetheless, if the merger does not lead to any efficiency gains (i.e., cost synergies), only under very restrictive conditions, reallocation of assets through structural remedies may satisfy the criterion of consumer surplus (Vergé 2010).

With regard to the nature of the *ex ante* merger control, asymmetric information problems arise between the merging firms and the competition agencies. Firms know precisely the level of their own efficiency gains expected, while the competition agency may doubt the real level of expected synergies (adverse selection). The merger remedies may be used to signal the true type of efficiency gains (see, e.g., Dertwinkel-Kalt and Wey (2016)).

The strategic trade-off for the merging firms is the following. On one hand, the greater the synergies of a proposed merger, the more companies may value their assets, and therefore are reluctant to propose significant commitments. Thus, theoretically, efficiency gains reduce the size of the proposed commitments (Cosnita and Tropéano 2009; Bougette 2010).

On the other hand, this argument does not take into account the temporal dimension of the trade-off. The EC's investigations are costly for merging firms. In this sense, firms may judge that it is preferable to divest more assets ("overfixing" strategy) than what exactly would suffice to resolve the competition concern, in order to avoid a long, and therefore, costly control procedure. Thus, in spite of the high level of synergies generated, the merging companies may then be inclined to give up on strong commitments simply because they are delay averse.

Cosnita-Langlais and Tropéano (2012) analyze the link between the efficiency defense and the use of remedies. They model the quality of information held by the competition agency and show that it may be best for the agency to prohibit the efficiency defense when the information quality is poor.

## Merger Remedies Retrospective Studies

With regard to empirical applications dedicated to remedies, several levels of study in the literature may be specified.

First, empirical analysis has focused on specific merger cases. These studies aim to assess the effectiveness of the decision made by competition agencies. Structural remedies may have been used and could be evaluated as such. In most cases, a difference-in-differences approach is adopted. The chosen econometric method allows for estimating a counterfactual, namely, what would have occurred in the absence of the studied merger, or in this case, in the absence of remedies. For example, Tenn and Yun (2011) show that the structural remedies used in the 2006 *J&J–Pfizer* merger resulted in the return of the premerger situation.

Second, another approach to evaluating selected remedies is to build a database of merger decisions including the ones with remedies and to analyze the determinants of these remedies. Several studies on the EC's data have been released in this perspective (see, e.g., Bougette and Turolla (2008)). Some of them have studied the merger control process in general, while others focus exclusively on the remedial decisions. For instance, Duso et al. (2011) show that remedies may be more effective when anticompetitive concerns are not too harmful and when applied to the first rather than the second investigation phase.

Time actually plays a considerable role in companies' strategy to provide more or less strict merger remedies, when needed. By studying 254 cases of mergers from the EC over the period 1999–2010, Ormosi (2012) shows that companies that do not use an efficiency defense strategy are actually more likely to reach a deal relatively early. The experience of law firms hired for the case has a high impact on the probability of using efficiency defense, but only for European cabinets. Less restrictive remedies are more likely to be proposed by the parties at the outset of the proceedings.

Based on the EC's data, Garrod and Lyons (2016) also show that the probability of early settlement is increasing in delay costs of the merging parties, decreasing in the uncertainty associated with the complexity of the economic assessment, and decreasing in the case load of the EC when resources are plentiful.

Finally, studies and reports have been conducted and prepared by the agencies themselves in a self-evaluation exercise of their decisions. The most notable of these is the European Commission (2005) that pointed out a number of practical problems in terms of monitoring, divested asset selection, and potential buyers, among others.

## Cross-References

- ► Competition Policy: France
- ► Difference-in-Difference
- ► Merger Control

## References

Bougette P (2010) Preventing merger unilateral effects: a nash–cournot approach to asset divestitures. Res Econ 64:162–174

Bougette P, Turolla S (2008) Market structures, political surroundings, and merger remedies: an empirical investigation of the EC's decisions. Eur J Law Econ 25:125–150

Compte O, Jenny F, Rey P (2002) Capacity constraints, mergers and collusion. Eur Econ Rev 46:1–29

Cosnita A, Tropéano JP (2009) Negotiating remedies: revealing the merger efficiency gains. Int J Ind Organ 27:188–196

Cosnita-Langlais A, Tropéano JP (2012) Do remedies affect the efficiency defense? An optimal merger-control analysis. Int J Ind Organ 30:58–66

Dertwinkel-Kalt M, Wey C (2016) Structural remedies as a signaling device. Inf Econ Policy 35:1–6

Duso T, Gugler K, Yurtoglu BB (2011) How effective is European merger control? Eur Econ Rev 55: 980–1006

European Commission (2005) Merger remedies study. Technical report

Garrod L, Lyons BR (2016) Early settlement in European merger control. J Ind Econ 64:27–63

Motta M, Polo M, Vasconcelos H (2007) Merger remedies in the European Union: an overview. Antitrust Bull 52:603–631

Ormosi PL (2012) Claim efficiencies or offer remedies? An analysis of litigation strategies in EC mergers. Int J Ind Organ 30:578–592

Tenn S, Yun JM (2011) The success of divestitures in merger enforcement: evidence from the J&J-Pfizer transaction. Int J Ind Organ 29:273–282

Vergé T (2010) Horizontal mergers, structural remedies, and consumer welfare in a cournot oligopoly. J Ind Econ 58:723–741

# Merit Goods

Valérie Clément[1] and Nathalie Moureau[2,3]
[1]MRE, EA 7491, Université de Montpellier, Montpellier, France
[2]Université Paul-Valéry MONTPELLIER 3, Montpellier, France
[3]ARTdev, UMR 5281, Université Paul Valéry, Montpellier, France

Merit goods are a category of goods, introduced in the debate by Musgrave (1957), which individuals tend to under- or over-consume because their preferences are "irrational" or "defective." This leads individuals to make suboptimal choices, which are detrimental to their well-being. Now, if they exist, merit goods must be produced by the government that must so to speak force individuals to consume the correct amount of these goods. In other words, the government must behave paternalistically.

The concept of merit goods was a precursor to the debates on paternalism within welfare economics. In particular, the interpretation of the merit goods concept through the meta-preferences approach helps in legitimizing legal intervention and achieving a more efficient regulation.

When Musgrave introduced the term "merit goods" (originally called *merit wants*), it was in an attempt to create a normative definition for government functions. Nevertheless, only three of the functions he studied in his article have gone down in history: (i) the provision of public goods (*service branch*), (ii) the redistribution of income (*distribution branch*), and (iii) economic regulation (*stabilization branch*). Yet, in this groundbreaking article, Musgrave also mentioned another category of goods which he called merit wants. He was referring to goods which are subject to "transfers in kind" (e.g., social housing) and for which the regulator's preferences override individual choices (Musgrave 1957 p. 341).

In 1959, Musgrave returned to this concept of merit goods by explicitly linking it to the issue of consumer sovereignty. In some cases, when choices made by people on the markets do not lead to a situation that maximizes their well-being, the regulator intervenes in order to address the limitations of individual preferences and correct people's choices in their own best interest.

It is nevertheless in his 1987 Palgrave article that Musgrave strengthened the definition he had introduced 30 years earlier. He clarified two points in particular which attracted most comments since they were first published.

Firstly, Musgrave confirms his initial theoretical claim that the justification for government intervention through merit goods is distinct from that linked to market failures and redistribution. Indeed, while links between merit goods, public goods, and externalities may have caused some confusion in his initial papers (Head 1966, 1969; Ver Eecke 2001), the Palgrave article provides clarification. In no way should merit goods be confused with public goods or externalities. Whereas in the case of public goods there is a link between consumers' willingness to pay and consumption levels, this link is broken in the case of merit goods. Furthermore, merit goods refer to situations where people's choices are detrimental to their own well-being without third parties being involved, as is the case with externalities.

Secondly, at the heart of the definition of merit goods lies the fact that if choices are detrimental to individual, it is because their current preferences are defective. Thus, choices then expressed in the market no longer equate with welfare. These *individual failures* could justify government interventions (Jones and Cullis 2002).

The reasons why choices made on the market may lead to a suboptimal situation have been the subject of extensive debate. In the article he wrote for the Palgrave Dictionary, Musgrave takes the view that situations in which people voluntarily delegate their choice to a more informed party, in a principal agent relationship, do not relate to merit goods. However, in his early works, he did not take this stance and had in fact used education as a prime example of merit goods. Indeed, at first he considered that the reason why education was compulsory was because people were not able to forecast the profit they would earn of such an investment. He nevertheless changed his mind, stating that it was simply an information issue encountered by the individual which justified a delegation of choice to another better informed party (For sure, one must admit that the government is better informed; that is not at all accepted in the economic literature.) (Musgrave 1987; West and McKee 1983). Defining the concept of merit goods is rather about highlighting the inconsistency of the preference standard in order to form judgements on individual well-being. Hence, it seems that even when full information is available, wrong choices can be made and lead to a suboptimal situation for the individual.

By definition, merit goods infringe consumer sovereignty and for this reason were put aside the standard welfare economics framework as the golden standard for paternalism (McLure 1968). However, there have been attempts to model merit goods in the context of welfare economics (Pazner 1972; Roskamp 1975; Wenzel and Wiegard 1981; Salanié and Treich 2009). These attempts perhaps reflect the need to justify an extremely widespread regulatory practice. For example, OECD data shows that two-thirds of European government bodies expenditure are somehow justified in terms of merit goods (Fiorito and Kollintzas 2004) and cannot be explained by standard market failure arguments.

If current short-term preferences are disqualified, the question arises of how "authentic" preference could be defined and what it stands for. The theoretical issue underlying this question lies in the possibility of articulating merit goods with the classical liberal principle of normative individualism. Musgrave did not evade the issue. In some of his papers, he noted that there is an elite who is in a position to know people's "true preferences" or "authentic preferences" (Musgrave 1969); in other papers, he refers to collective norms or "community preferences" (Musgrave 1987).

Another way to justify the concept of merit good in the economic framework, which seems more in line with the theoretical issue, involved expanding the area of individual preferences beyond market preferences, displayed through the willingness to pay and choice, by introducing the notions of "multiple-selves" and "metapreference." The economic agent is then defined by a collection of different and independent personalities (Harsanyi 1955; Elster 1979; Etzioni 1986), each of which leads to a separate classification of available options. The individual is no longer a unified person and may struggle to control his/her behavior (Schelling 1984). Equally, the individual may have the ability to assess and reflect on his/her own tastes and preferences that are expressed through second-order preferences or meta-preferences (Frankfurt 1971; Jeffrey 1974; Sen 1977; Hirschman 1984; George 1998). These help reflect the individual's dissatisfaction with a choice that he/she nevertheless made. The regulator then appears as a mediator between preferences displayed on the market on the one hand and reflexive preferences on the other, this mediation being then carried out under merit goods (Brennan and Lomasky 1983).

Incorporating into the analysis this idea of the existence of several ranges of preferences enables decisions made by policymakers, legislators, and judges to be perceived as an expression of second-order or meta-preferences, which, in turn, allows for the regulation implemented under merit goods to lead to higher efficiency than that implemented by the market, while respecting the individualistic foundations of collective choice. In this sense, the interpretation of Musgrave's concept through reflexive preferences is particularly relevant when analyzing economic policies and regulation policies within a law and economics approach. As noted by Kirchgässner (2017), this ties in with an important tradition in political philosophy which, from Buchanan and Tullock (1962) to Rawls (1971), combines the choice of a constitution or of the general principles on which society is organized with higher-order preferences.

With the developments of libertarian paternalism (Sunstein and Thaler 2003), the question of the role played by merit goods in the economic framework remains topical. By exploring faulty reasoning and rationality defaults, behavioral economics actually deepen the empirical content of Musgrave merit goods' rationale. Nevertheless, this strand of literature quite surprisingly did not refer to the concept of merit good in its developments of a new framework for the state regulation. Obviously, behavioral economists argue against merit goods' intervention as it represents hard paternalism restricting individual choices through prohibition or taxation. Behavioral economists favor nudge practices, where the regulator helps people to make the best choice through a change in the frame or in the environment of choices (Thaler and Sunstein 2008). In any case, there should not be any restriction of the available options provided through the market allocation.

Then, behavioral economics did not pay heed to the contribution of the merit goods' argument to the normative justification for the government intervention. In this respect, the continuity with the pioneering concept of Musgrave is certainly to be found in the role and definition devoted to the merit goods in law and economics following Calabresi (2016). Calabresi complements and extends the definition of Musgrave adding two reasons why merit goods should not be (and actually are not) allocated through markets: the refusal of the "commodification" and pricing of certain goods and the refusal of an allocation based on people's willingness to pay given the vast inequalities in wealth distribution in our societies.

In the first category, people object to the use of monetary evaluation and measurement for being conducive toward unacceptable trade-offs, for example, trade-offs implying life or safety and money. The second category of merit goods following Calabresi includes those goods whose measurement in monetary term is no longer objectionable, but people oppose the use of the pure market mechanisms because the allocation thus depends on the prevailing unequal distribution of wealth; examples are military service or the right to obtain body parts (blood, kidney, etc.) or the right to a basic education. Taking seriously people's actual preferences embedded in these two merit goods categories allows the society to avoid "indirect external moral costs," Calabresi argues, that arise from the denial of people's objection to commodification and to the neglecting of the distributional consequences of the pure market allocation. Hence the allocation of merit goods should rest on hybrid mechanisms involving either modified market or modified command schemes if people preferences for merit goods are taken into account. Tort laws provide a prominent example of such hybrid mechanisms in the case of objection to commodification that lead to lessen the externalities created by merit goods, that is, “moral costs” people would bear were the merit goods (life and safety) priced directly through the market, Calabresi points out.

Merit goods lie at the heart of law and economics as Calabresi conceived it, first of all because the inclusion of people’ preferences about commodification and equality enables regulation policies to be efficient in the sense that third-party moral costs are fully integrated and lastly because the thorough study of the law and the legal institutions should serve to identify merit goods and to elicit people’s preferences about merit goods. This renewal of the merit goods’ argument confirms the initial statement of Musgrave that merit goods were a category of goods that called for the expansion of the standard economic model.

## Cross-References

- ▶ Libertarian Paternalism
- ▶ Nudge

## References

Brennan G, Lomasky L (1983) Institutional aspects of ‘merit goods’ analysis. Finanzarchiv 41(2):183–206

Buchanan JM, Tullock G (1962) The calculus of consent: logical foundations of constitutional democracy. University of Michigan Press, Ann Arbor

Calabresi G (2016) The future of law and economics. Yale University Press, New Haven

Elster J (1979) Ulysses and the sirens: studies in rationality and irrationality. Cambridge University Press, New York

Etzioni A (1986) The case for a multiple utility conception. Econ Philos 2(2):159–183

Fiorito R, Kollintzas T (2004) Public goods merit goods and the relation between private and government consumption. Eur Econ Rev 48:1367–1398

Frankfurt HG (1971) Freedom of the will and the concept of a person. J Philos 68(1):5–20

George D (1998) Coping rationally with unpreferred preferences. East Econ J 24(2):181–194

Harsanyi J (1955) Cardinal welfare individualistic ethics and interpersonal comparisons of utility. J Polit Econ 63(4):309–321

Head JG (1966) On merit goods. Finanzarchiv 25(1):1–29

Head JG (1969) Merit goods revisited. Finanzarchiv 28(2):214–225

Hirschman AO (1984) Against parsimony: three easy ways of complicating some categories of economic discourse. Am Econ Rev 74(2):89–96

Jeffrey R (1974) Preferences among preferences. J Philos 71(13):377–391

Jones P, Cullis J (2002) Merit want status and motivation: the knight meets the selfloving butcher brewer and baker. Public Financ Rev 30(2):83–101

Kirchgässner G (2017) Soft paternalism merit goods and normative individualism. Eur J Law Econ 43(1): 125–152

McLure CE (1968) Merit wants: a normative empty box. Finanzarchiv 27(2):474–483

Musgrave RA (1957) A multiple theory of budget determination. Finanzarchiv 17(3):333–343

Musgrave RA (1959) The theory of public finance: a study in public economic. McGraw-Hill Book Company, New York

Musgrave RA (1969) Provision for social goods. In: Margolis J, Guitton H (eds) Public economics. McMillan, London, pp 124–144

Musgrave RA (1987) Merit goods. In: Eatwell J, Milgate M, Neuman P (eds) The New Palgrave: a dictionary of economics. Macmillan, London, pp 452–453

Pazner EA (1972) Merit wants and the theory of taxation. Public Financ 27(4):460–472

Rawls J (1971) A theory of justice. Harvard University Press, Cambridge, MA

Roskamp KW (1975) Public goods merit goods private goods Pareto optimum and social optimum. Public Financ 30(1):61–69

Salanié F, Treich N (2009) Regulation in Happyville. Econ J 119(537):665–679

Schelling TC (1984) Self-command in practice in policy and in a theory of rational choice. Am Econ Rev 74(2):1–11

Sen AK (1977) Rational fools: a critique of the behavioural foundations of economic theory. Philos Public Aff 6(4):317–344

Sunstein CR, Thaler RH (2003) Libertarian paternalism is not an oxymoron. University Chicago Law Rev 70(4): 1159–1202

Thaler RH, Sunstein CR (2008) Nudge: improving decisions about health, wealth, and happiness. Yale University Press, New Heaven

Ver Eecke W (2001) The concept of a merit good: the ethical dimension in economic theory and the history of economic thought or the transformation of economics into socio economics. J Socio-Econ 27(1):133–153

Wenzel HD, Wiegard W (1981) Merit goods and second-best taxation. Public Financ 36(1): 125–139

West EG, MacKee M (1983) De gustibus Est disputandum the phenomenon of merit wants revisited. Am Econ Rev 7(5):1110–1121

## Method of Merchants

▶ Lex Mercatoria

## Mises, Ludwig von

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

**Abstract**

Ludwig Heinrich Edler von Mises (September 29, 1881, in Lemberg – October 10, 1973, in New York City) was a classical liberal philosopher, sociologist, and one of the most influential adherents to the Austrian school of economics. He made major contributions to the epistemology of the social sciences and to many areas of general economics, especially the fields of value theory, monetary theory, and business cycle theory. In his habilitation thesis of 1912, *The Theory of Money and Credit*, he laid down the foundations of what would later become Austrian business cycle theory. The whole body of Misesian economics which is based on praxeology, the rational investigation of human decision making, is comprised in his 1949 *Human Action: A Treatise on Economics*. Among his disciples were such noteworthy economists as Friedrich August von Hayek (1899–1992) and Murray N. Rothbard (1926–1995).

## Life, Work, and Influence of Ludwig von Mises

### Life, Family, and Personal Background

Ludwig von Mises (Fig. 1), son of Arthur Edler von Mises and Adele Landau, was born in the city of Lemberg, Galicia, in the former Austro-Hungarian Empire (today Lviv, Ukraine) on September 29, 1881. He is the older brother of the famous Harvard mathematician Richard Martin von Mises (1883–1953). His youngest brother Karl died of scarlet fever at the age of 12 in 1903.

**Mises, Ludwig von, Fig. 1** Ludwig von Mises

In 1881 Emperor Franz Joseph granted Ludwig's great grandfather Meyer Rachmiel Mises a patent of nobility and the right for him and his lawful offspring to bear the honorific title "Edler" (Hülsmann 2007, p. 15). Ludwig then was the first member of his family to be born a nobleman. The whole Mises family was heavily involved in the construction and financing of Galician railways after the 1840s. So it came to pass that Ludwig's father was a construction engineer for the Czernowitz railway company in Lemberg before moving to Vienna no later than 1891 (Hülsmann 2007, p. 21).

Ludwig entered the *Akademische Gymnasium* of Vienna in 1892. There, he was given an education stressing among other things the classical languages, Latin and ancient Greek. Reading Virgil, he found a verse that he chose to be his motto for life: "Tu ne cede malis sed contra audentior ito" (translated: Do not give in to evil, but proceed ever more boldly against it) (Mises 2009, p. 55). Much later he would point out the importance of the classical literature, and in particular of the ancient Greeks, for the emergence of liberal social philosophy in *The Anti-Capitalistic Mentality*

(Mises 1956; Hülsmann 2007, p. 34). After having completed the Gymnasium in May 1900, when his major interests were in politics and history, he enrolled in the Department of Law and Government Science at the University of Vienna.

Initially, Mises studied under the Marxist sociologist and economist Carl Grünberg (1861–1940) who was an adherent of the German historical school. At the age of only 20, Mises published his first scholarly work in one of Grünberg's books on the peasants' liberation and agrarian reforms in the Bukovina (Hülsmann 2007, p. 68). He graduated with the first *Staatsexamen* in 1902 in history of law from the University of Vienna. In 1906, after completion of the military service, he passed the second and third *Staatsexamina* in law and government science. Thereafter, he was awarded a doctorate degree, *doctor juris utriusque*, from the same university by passing his juridical, political science, and general law exams.

Mises in the course of his studies shifted gradually away from the influence of historicism which was at the time the predominant method in the social sciences in the German-speaking world. Highly influenced by economists Carl Menger (1840–1921) and his follower Eugen Böhm von Bawerk (1851–1914), he published his habilitation thesis in 1912 under the title *Theorie des Geldes und der Umlaufsmittel* (translated in 1934 as *The Theory of Money and Credit*). In fact, it was Menger's *Grundsätze der Volkswirtschaftslehre* (*Principles of economics*, Menger 1871) and Böhm-Bawerk's *Kapital und Kapitalzins* (*Capital and Interest*, von Böhm-Bawerk 1890) originally published in 1884, two foundational works of the Austrian school of economics, that had a tremendous impact on Mises's thinking. Mises wrote in his *Memoirs* that it was through Menger's book that he became an economist (Mises 2009, p. 25). After his habilitation, he worked as a civil servant for the Austrian chamber of commerce and became a *Privatdozent*, an unsalaried lecturer, at the University of Vienna – the highest academic rank he would ever achieve in his native Austria.

During the First World War, Mises served as an artillery officer in the Austro-Hungarian army and spent several months at the front. It was only in December 1917 that he was ordered to join the department of war economy in the War Ministry of Vienna. He resumed his teaching activities at the University in the spring of 1918. Among his students at that time was Richard von Strigl (1891–1942) who became an important Austrian economist of the interwar period. In 1919, he finished a manuscript under the working title *Imperialismus* which would contain his analysis of the causes of the Great War and the political challenges for postwar Europe as well as his personal war experiences. The book was eventually published under the title *Nation, Staat und Wirtschaft* (translated in 1983 as *Nation, State and Economy*, (von Mises 1919). This book, although less well known today, established Mises as the foremost champion of classical liberalism in the German-speaking world and eventually in all of Europe (Hülsmann 2007, p. 300).

Mises made a new personal acquaintance with the famous Max Weber (1864–1920), who came to teach at the University of Vienna in September of 1917. Their professional relationship was characterized by mutual respect and admiration. Weber had a remarkable impact on Mises's writings on the methodological and epistemological problems of the social sciences in the 1920s. On the other hand, Weber praised Mises's monetary theory as "the most acceptable" of the time (Hülsmann 2007, p. 288).

Only a few years later, in 1922, Mises published "the most devastating analysis of socialism yet penned" (Hazlitt 1956, p. 35), *Die Gemeinwirtschaft* (translated in 1936 as *Socialism* (von Mises 1951), a book in which he gives a detailed and thorough explanation of the problem of economic calculation under socialism. In 1927, his *Liberalismus* (translated in 1962 as *Liberalism* (von Mises 1985) was first published). Two years later, *Kritik des Interventionismus* (translated in 1977 as *A Critique of Interventionism*, von Mises 1996) appeared, an anthology of articles he wrote in the early 1920s.

Throughout the 1920s until 1934, Mises held a weekly private seminar in his office at the chamber of commerce to which students and scholars not only from Vienna but from all over Europe and the whole world were attracted. Among the numerous participants were Austrian economist Gottfried Haberler (1900–1995), Frenchman Francois Perroux (1903–1987), American economist Frank Knight (1885–1972), and even four Japanese economists (Itschitani, Midutani, Otaka, Takemura) (Hülsmann 2007, p. 674). However, as one of the best-known students and colleagues of Mises, one has to mention Friedrich August von Hayek who attended the seminar on a regular basis. Hayek became the director of the *Austrian Institute for Business Cycle Research* that Mises had established (Hülsmann 2007, p. 454). Later, Hayek would take a position at the *London School of Economics* which was an important step in his academic career.

The political developments made life in Vienna increasingly unpleasant for Mises. Being of Jewish descent and an ardent critic of socialism, he had to flee Vienna in 1938 after occupation and annexation of Austria by the German national socialists. With great foresight he had already accepted a position at the Geneva-based *Institut de Hautes Études Internationales* (Graduate Institute of International Studies) in 1934, where he became visiting professor. There, under the company of William E. Rappard (1883–1958) and Paul Mantoux (1877–1956) who led the school together for about 20 years, he had some of the most productive and fruitful years of his scholarly career, culminating in the publication of his opus magnum *Nationalökonomie* in 1940. In 1938, he would marry his longtime companion Margit Serény in Geneva.

As history moved on, Geneva as well seemed no longer to be a safe place for Mises and his wife, due to the rising threat of national socialism and ultimately the outbreak of World War II. Mises was high on the list of wanted men. In March 1938, unidentified men broke into Mises's apartment in Vienna. A few days later, the Gestapo confiscated his books, correspondences, personal records, as well as other belongings. At the end of the war, the Red Army found his files in a train in Bohemia. They were brought to a secret archive in Moscow, where they would be rediscovered in 1991, 18 years after his death (Ebeling 1997). Mises never saw his documents again. He thought that they had been destroyed during the war.

Ludwig and Margit von Mises decided to flee Europe. After a long and arduous trip, they arrived at the docks of New York City on the third of August, 1940. At almost 60 years of age, Mises had to start a new life. He quickly sought contact with potential supporters. His brother was already professor at Harvard University. Henry Hazlitt (1894–1993), journalist at the *New York Times* and a great admirer of Mises, turned out to be very helpful. With the financial support of the *William Volker Fund,* he arranged a visiting professorship for Mises at *New York University* (NYU). Mises would remain a visiting professor for more than 20 years until he retired in the Spring of 1969, as the oldest active professor in the United States (Rothbard 1988, p. 44). His salary during that period would always be paid from private funds.

Mises set up a seminar at NYU in the spirit of his Vienna private seminar, where he gathered a diverse group of journalists, businessmen, scholars, and young university and even high school students. Hans Sennholz (1922–2007) and Louis Spadaro (1913–2008) were his first doctoral students in the United States. Other seminar members were William Peterson (1921–2012), George Reisman (born 1937), his classmate Ralph Raico (born 1936), Israel Kirzner (born 1930), and most notably Murray N. Rothbard. According to his wife, Mises encouraged all of his students to pursue scholarly work, hopeful that one of them might develop into a second Hayek (Mises 1976, p. 135). He might have found his "second Hayek" in Rothbard, although both Hayek and Rothbard developed Misesian ideas in quite different directions.

Mises's first books written in English appeared in 1944, *Omnipotent Government* (von Mises 1944a) and *Bureaucracy* (von Mises 1944b), both published by Yale University Press. *Human Action*, the extended English edition of his opus magnum *Nationalökonomie*, was published in 1949. It contains the culmination of Misesian economic theory. Among several other books and numerous articles, he published his last great work in 1957, *Theory and History* (von Mises

1957), a philosophical treatise that builds a bridge between economic theory and human history and explains the true relation between those two disciplines (Rothbard 1988, p. 110). Having devoted his whole life to the enhancement of economic theory and the social sciences, as well as the promotion of peace and individual liberty, Ludwig von Mises died at the age of 92 on October 10, 1973, in New York City.

## Work and Influence

When Ludwig von Mises entered university in 1900, the social sciences in the German-speaking world, including economics, were predominantly influenced by historicism – ideologically a forerunner of positivism. Gustav Schmoller (1838–1917), under whom Mises's first university teacher Carl Grünberg studied, was the leading intellectual of the German historical school. Mises recognized fundamental flaws within this school of thought and the academic tendencies of his time, namely, the state orientation:

> It was my intense interest in historical knowledge that enabled me to perceive readily the inadequacy of German historicism. It did not deal with scientific problems, but with the glorification and justification of Prussian policies and Prussian authoritarian government. The German universities were state institutions and the instructors were civil servants. The professors were aware of this civil-service status, that is, they saw themselves as servants of the Prussian king. (Mises 2009, p. 7, as cited in Rothbard 1988, p. 51)

Mises was interested in positive social sciences, not in the promotion of political measures by the government. He was convinced that there are absolutely and uncompromisingly, in his words "apodictically," true statements in the realm of social sciences that are not mere tautologies or conventions – an idea that historicists and positivists would decidedly reject. Mises on the other hand rejected the relativism of the historicists. He first found such uncompromisingly true statements in Menger's *Grundsätze der Volkswirtschaftslehre*, which is widely considered to be the foundational work of the Austrian school of economics. This book made Mises an economist and an "Austrian." Mises, among the other Austrian economists of the second and third generation after Menger, dwelled on this epistemological and methodological point the most vehemently. He considered economics to be an "a priori" science, rooted in the broader discipline of the rational investigation of human behavior and decision making that he would term praxeology.

The Misesian body of economic theory is derived from the undeniable fact that human beings exist and act, that is, they are consciously employing means to attain ends. One cannot argue that human beings do not act, that they do not employ means to attain ends, since such an argument would precisely be an action as described above. By acting, human beings make choices, they demonstrate preferences, and they engage in exchanges. They make value judgments – they decide for one opportunity and they forgo others, that is, they pay a price. Mises shows that all fundamental economic concepts, such as value, exchange, preferences, prices, costs, and gains, are inherent to human action:

> Action is an attempt to substitute a more satisfactory state of affairs for a less satisfactory one. We call such a willfully induced alteration an exchange. A less desirable condition is bartered for a more desirable. What gratifies less is abandoned in order to attain something that pleases more. That which is abandoned is called the price paid for the attainment of the end sought. The value of the price paid is called costs. Costs are equal to the value attached to the satisfaction which one must forego in order to attain the end aimed at.
>
> The difference between the value of the price paid (the costs incurred) and that of the goal attained is called gain or profit or net yield. Profit in this primary sense is purely subjective, it is an increase in the acting man's happiness, it is a psychical phenomenon that can be neither measured nor weighed. There is a more and a less in the removal of uneasiness felt; but how much one satisfaction surpasses another one can only be felt; it cannot be established and determined in an objective way. A judgment of value does not measure, it arranges in a scale of degrees, it grades. It is expressive of an order of preference and sequence, but not expressive of measure and weight. Only the ordinal numbers can be applied to it, but not the cardinal numbers. (Mises 1998, p. 97)

As one can see from the above quote, Mises's value theory is completely subjective, with the consequence that he would not only refuse the

Marxian labor theory of value but also the welfare economics of a Léon Walras (1834–1910). For Mises, value or utility is something that only the individual attaches to a good. It cannot be measured and therefore interpersonal utility comparisons and concepts like "aggregate welfare" are invalid. On this issue, Mises takes the same line as Czech economist Franz Cuhel (1862–1914) (Rothbard 1988, pp. 19 and 59). The only common denominator of individual preferences and evaluations, through which heterogeneous goods and services can be compared somewhat objectively, is the price system of a free market. However, this is not to be confused with measuring utility in terms of money prices.

From this insight, Mises derives his critique of central planning. In a centrally planned economy, where the means of production are not privately owned and therefore are not sold and bought on the market, there are no market prices for production or capital goods. Consequently, there is no way for the central planners to determine which production processes or which combinations of capital resources, for the production of some desired good, are economically efficient and in line with consumer preferences. In short, economic calculation is impossible under socialism. *Socialism: An Economic and Sociological Analysis*, the relevant book for this topic, is to a large extent built on an earlier article that started the socialist calculation debate, *Economic Calculation in the Socialist Commonwealth*. According to Friedrich August von Hayek, this debate and in particular Mises's contribution had a remarkable impact on the social scientists and economists of his generation, who predominantly favored socialism over the free market (Hayek 1981).

In *A Critique of Interventionism*, Mises provides us with a very distinctive clarification of another crucial question concerning political economy. If socialism is inherently dysfunctional, what about a middle-of-the-road solution – a mixed economy that is neither a complete free market system nor full-blown socialism? Can we combine the benefits of both systems in a suitable way? Mises's answer is no. Such a system of interventionism would be inherently unstable. For every problem that the government detects and attempts to solve through intervention into the economy, it will most of the time not only not solve the problem but also create new ones. After each intervention the government can then either take the initial intervention back or go on to intervene further into the economy. Mises argued that interventionism is a slippery slope and ultimately leads to socialism. The only alternative is provided by a genuine free market system (Bagus 2013). This idea of the self-reinforcing character of interventionism has been picked up by Hayek in his best-selling popular book *The Road to Serfdom* (von Hayek 2005). Hayek does not deny the impact that Mises had on his political philosophy, although some economists and philosophers who see themselves more in line with the Misesian tradition and the Rothbardian interpretation thereof, such as Walter Block, criticized Hayek sharply for his "lukewarm" defense of laissez-faire capitalism (Block 1996).

Although his critique of the state is extensive, Mises would not deny the necessity of the state altogether. As unmistakably explained in *Liberalism*, he assigns a unique task to the state, namely, the enforcement of law and in particular the protection of private property. Mises thereby remains in the tradition of classical liberalism (Hoppe 2013). It was Murray N. Rothbard, certainly Mises's most influential American student, who would enrich the Misesian analysis of the state with his theory of rational ethics developed in his philosophical treatise *The Ethics of Liberty* (Rothbard 2003) and push it to its ultimate conclusion: the state should be abolished. Rothbard spearheaded the anarcho-capitalistic movement in the United States that today is more vibrant than ever.

When it comes to pure economics, Mises probably made his most important contributions in the area of monetary theory and business cycle theory. In his habilitation thesis, *The Theory of Money and Credit* (von Mises 1912), he elaborates on Menger's account of the origins of money out of barter trade to solve a challenge that has been laid down to the Austrian economists by Karl Helfferich (1872–1924) in 1903. In his work *Money*, Helfferich correctly pointed out that the "Austrians," despite their comprehensive microeconomic analysis of markets and prices for goods and services, had not yet managed to solve the problem of money. Money had been treated separately from the rest of economics in a "macro-box," independently of utility, value, and relative prices (Rothbard 1988, p. 55). When trying to explain the value of money, one ended up in a circular argument. Money is a special good that is demanded not for consumption but for exchange – in the present or some point in the future. It is demanded because it has an exchange value, but it has its exchange value only because it is demanded. This circular reasoning posed a problem that Mises was able to solve by incorporating the time dimension into this interdependency (Hülsmann 2013).

The demand for money today is determined by the exchange value of money in previous periods. The demand for money in previous periods was determined by its exchange value in still earlier periods, and so the chain of reasoning goes backward in time. Do we end up in an infinite regress? No, because at some point back in time, the money good has been demanded, as any other good, for its value in consumption. At this point, we are back in a barter economy. This is the starting point. In a transition from direct to indirect exchange, in order to overcome the problem of the double coincidence of wants, some goods and eventually one single good will become universally accepted as a medium of exchange. This good will become money. Traditionally, precious metals such as gold and silver have played this role, since they are rare, homogeneous, highly divisible, transportable, and durable. Mises's explanation, which we call the *Regression Theorem*, "was a remarkable achievement, because for the first time, the micro/macro split that had begun in English classical economics with Ricardo was now healed" (Rothbard 1988, p. 57). Money was incorporated into the rest of economic theory and had no longer to be treated separately. For a more detailed outline and interpretation of Mises's monetary theory, the interested reader may have a look at *Theory of Money and Fiduciary Media* edited by Guido Hülsmann (2012), an anthology of essays in celebration of the centennial of Mises's major work.

The next fundamental contribution by Ludwig von Mises is his theory of business cycles that emerged out of his monetary theory. Thinking about the way money evolves on the market according to Menger and Mises, one recognizes that our current monetary system is quite special, in the sense that our money is not backed by any real good. We are living in a fiat money system. Our money is money by government decree. However, this regime has been transformed into what it is today by government intervention out of a gold-backed monetary system, and its mere existence does therefore not disprove the Mengerian-Misesian account. That it is advantageous for any government to possess the monopoly over the production of an unbacked currency is quite obvious. It enables the government to create money out of thin air through credit expansion and to lower interest rates artificially. The Keynesian view holds that this is a necessary political tool for anti-cyclical economic stabilization and therefore the cure of the business cycle which is inherent to free market capitalism. Mises, on the contrary, sees the root cause of the business cycle not in the free market itself but precisely in artificial credit expansion. Mises's explanation is as follows.

He built his theory out of three preexisting components, the business cycle model of the Currency School, the differentiation between the "natural rate of interest" and the "bank rate of interest" by Swedish economist Knut Wicksell (1898), and the Böhm-Bawerckian capital and interest theory (Rothbard 1988, p. 63; Hülsmann 2007, Chapter 6). Mises argues that pumping money into the economy by expanding credit and lowering interest rates under the "natural" time preference level cause excess malinvestments in capital goods industries.

In Mises's view, the interest rate is not an arbitrary number that should be interfered with. Instead, it is the price that tends to accommodate the roundaboutness of production processes or investment projects to the available subsistence fund in the economy. Usually, interest rates fall, when consumers save more and thereby increase the subsistence fund. However, in the case of artificial credit expansion, the decrease in interest rates and the subsequent excess investments are not justified or covered by real savings. Therefore, some of these investments, sooner or later, have to be liquidated, when it turns out that the subsistence fund in the economy is too small to finish them all.

Note that artificial credit expansion is not a purely political phenomenon that can only be brought about by the state. It can and did happen on the market, on behalf private banks. However, it is only through the institutionalization of credit expansion, by establishing a central bank system, that it has reached today's magnitude.

Hayek won the Nobel Memorial Prize in economics in 1974, the year after Mises's death. He received it precisely for his contributions to business cycle theory, that is, the work he did in the 1920s and 1930s as an "ardent Misesian" (Rothbard 1988, p. 112). This decision by the Nobel Prize committee gave a boost to Austrian economics in general and Austrian business cycle theory in particular. According to Rothbard, it marked a "revival of Austrian economics."

The financial crisis of 2007 has attracted more attention to the Austrian explanation of booms and busts. Subsequent to this event, Ron Paul, presidential candidate in the United States, gained increased recognition for his critique of the Federal Reserve System and his call to abolish it. His arguments are essentially backed by Misesian ideas.

Mises's "last great work" and "by far the most neglected" was *Theory and History.* Yet, it "provides the philosophical backstop and elaboration of the philosophy underlying *Human Action*" (Rothbard 1985). He had a lot of opponents concerning his political recommendations, but it was his methodological uniqueness that made him an academic outlier for his whole life. In *Theory and History,* Mises provides the philosophical justification for considering the social sciences, including economics, as a completely different branch than the natural sciences. The subjects of investigation in the social sciences are human beings, with minds, who have preferences and make choices – who have goals and try to attain these goals. They act purposefully and they change their minds constantly. Human action does not follow mechanical and quantifiable laws like atoms or molecules do in physics. The empirical approach to the social sciences is ignorant of the uniqueness and the individuality of human decisions and the environment in which each decision is made. Human action is not replicable. By pointing out this insight, Mises made the case for methodological dualism. In the natural sciences, the researcher looks at data of repeated identical experiments in which all relevant variables can be controlled to find and isolate causal relationships. In economics, we already know the ultimate cause: humans act to attain their ends. This knowledge is not hypothetical. It is in Mises's words apodictically true and is not subject to empirical falsification – and neither are logical deductions from this primordial fact. Mises has been called unscientific and mystical for calling economics an "a priori" science, in the same way as mathematics is an "a priori" science. He has been criticized for being ignorant of economic history. His rebuttal is contained in *Theory and History.* It is the historicist-positivist-empiricist economist who overlooks the unique character of historical events by trying to draw generalized conclusions from observable data and thereby ignores the single common feature of all economic data: purposefully acting individual human beings. For Mises, it is not history that can provide us with a reliable theory. It is only by means of a theoretical framework that we can make sense out of historical data.

Today, the ideas of Ludwig von Mises and his intellectual followers are promoted more effectively than ever before by the *Ludwig von Mises Institute* in Auburn, Alabama. It was founded in 1982 by Llewellyn H. Rockwell Jr., Burton Blumert, and Murray N. Rothbard. Its website, www.mises.org, makes a large number of books, journal articles, and other writings available for free. It is worth a look for every interested reader. There exist a number of professional journals and periodicals published by the institute, including *The Journal of Libertarian Studies* (1977–2008), *The Review of Austrian Economics* (1987–1998), and *The Quarterly Journal of Austrian Economics* (since 1998).

Today, without formal ties among them, there exist more than 14 Mises Institutes worldwide, including those in Belgium, Switzerland, Germany, Portugal, Sweden, Finland, Czech Republic, Romania, Poland, Russia, Canada, Brazil, Ecuador, and Japan. There also exist a Spanish Language Institute and a Mises Institute Europe.

## References

Bagus P (2013) Mises' Staats- und Interventionismuskritik, published in Ludwig von Mises – Leben und Werk für Einsteiger, Finanzbuchverlag

Block W (1996) Hayek's road to Serfdom. J Libert Stud 12(2):339–365. Available online http://mises.org/journals/jls/12_2/12_2_6.pdf

Ebeling R (1997) Mission to moscow: the lost papers of Ludwig von Mises. Liberty Magazine 10(5). A presentation by R. Ebeling at Universidad Francisco Marroquín is available online: http://newmedia.ufm.edu/gsm/index.php/Mission_to_Moscow:_Discovering_the_%22Lost_Papers%22_of_Ludwig_von_Mises,_and_their_significance

Hazlitt H (1956) Two of Ludwig von Mises' most important works. In: Sennholz M (ed) On freedom and free enterprise – essays in honor of Ludwig von Mises. D. Van Nostrand. Available online http://mises.org/document/3327/On-Freedom-and-Free-Enterprise-Essays-in-Honor-of-Ludwig-von-Mises

Helfferich K (1903) Das Geld, 1th edn. Hirschfeld, Leipzig

Hoppe H-H (2013) Ludwig von Mises und der Liberalismus, republished in Ludwig von Mises – Leben und Werk für Einsteiger, Finanzbuchverlag

Hülsmann J-G (2007) Mises – the last knight of liberalism. Ludwig von Mises Institute, Alabama. Available online http://mises.org/document/3295/Mises-The-Last-Knight-of-Liberalism; for a presentation by the author see http://www.youtube.com/watch?v=h9BgKL5kX4U

Hülsmann J-G (ed) (2012) Theory of money and fiduciary media. Ludwig von Mises Institute, Alabama. Available online http://mises.org/document/7018/Theory-of-Money-and-Fiduciary-Media

Hülsmann J-G (2013) Mises' Geldtheorie, published in Ludwig von Mises – Leben und Werk für Einsteiger, Finanzbuchverlag

Menger C (1871) Grundsätze der Volkswirtschaftslehre, Wilhelm Braumüller – k.k. Hof- und Universitätsbuchhändler. Available online http://docs.mises.de/Menger/Menger_Grundsaetze.pdf; for the English translation see http://mises.org/document/595/Principles-of-Economics

Rothbard MN (1985) Preface to theory and history by Ludwig von Mises. Ludwig von Mises Institute, Alabama

Rothbard MN (1988) The essential von Mises. Ludwig von Mises Institute, Alabama. Available online http://mises.org/document/3081/The-Essential-von-Mises

Rothbard MN (2003) The ethics of liberty. New York University Press, New York and London

von Böhm-Bawerk E (1890) Capital and interest. Macmillan, London/New York. Available online http://mises.org/document/164/Capital-and-Interest; for the German original see https://archive.org/details/kapitalundkapita01bh

von Hayek FA (1981) Foreword to socialism: an economic and sociological analysis. Liberty Fund, Indianapolis

von Hayek FA (2005) The road to Serfdom with the intellectuals and socialism. Institute for Economic Affairs. Available online http://mises.org/document/2402/The-Road-to-Serfdom

von Mises L (1912) Theorie des Geldes und der Umlaufsmittel. Verlag von Duncker und Humblot, München/Leipzig. Available online http://mises.org/document/3298/Theorie-des-geldes-und-der-Umlaufsmittel; for the English edition see http://mises.org/document/194/The-Theory-of-Money-and-Credit

von Mises L (1919) Nation, Staat und Wirtschaft. Beiträge zur Politik und Geschichte der Zeit. Manzsche Verlags- und Universitäts-Buchhandlung, Vienna/Leipzig; eager as nation, state, and economy (trans: Leland B). New York University Press, New York (1983). Available online http://mises.org/document/1085/Nation-State-and-Economy

von Mises L (1944a) Bureaucracy. Yale University Press, New Haven. Available online http://mises.org/document/875/Bureaucracy

von Mises L (1944b) Omnipotent government: the rise of the total state and total war. Yale University Press, New Haven. Available online http://mises.org/document/5829/Omnipotent-Government-The-Rise-of-the-Total-State-and-Total-War

von Mises L (1951) Socialism – an economic and sociological analysis. Yale University Press, New Haven. Available online http://mises.org/document/2736/Socialism-An-Economic-and-Sociological-Analysis; for the German original see http://mises.org/document/3297/Die-Gemeinwirtschaft

von Mises L (1956) The anti-capitalistic mentality. Van Nostrand, Princeton. Available online http://mises.org/document/1164/The-AntiCapitalistic-Mentality

von Mises L (1957) Theory and history – an interpretation of social and economic evolution. Yale University Press, New Haven. Available online http://mises.org/document/118/Theory-and-History-An-Interpretation-of-Social-and-Economic-Evolution

von Mises M (1976) My years with Ludwig von Mises. Ludwig von Mises Institute, Alabama. Available online http://mises.org/document/3199/My-Years-with-Ludwig-von-Mises

von Mises L (1985) Liberalism: in the classical tradition. The Foundation for Economic Education, Inc. Irvington-on-Hudson, New York 10533. Available online http://mises.org/document/1086/Liberalism-In-the-Classical-Tradition; for the German original see http://mises.org/document/5452/Liberalismus

von Mises, L (1996) A critique of interventionism, irvington-on-hudson. Foundation for Economic Education, New York. Available online http://mises.org/document/877/Critique-of-Interventionism-A

von Mises L (1998) Human action (the Scholar's edition). Ludwig von Mises Institute, Alabama. Available online http://mises.org/document/3250/Human-Action; for the German predecessor *Nationalökonomie* see http://mises.org/document/5371/Nationalokonomie-Theorie-des-Handelns-und-Wirtschaftens
von Mises L (2009) Memoirs. Ludwig von Mises Institute, Alabama. Available online http://mises.org/document/4249/Memoirs
Wicksell K (1898) Geldzins und Güterpreise. Gustav Fischer Verlag, Jena. The English version is available online http://mises.org/document/3124/Interest-and-Prices

# Money Laundering

Donato Masciandaro
Department of Economics, Bocconi University, Milan, Italy

## Abstract

Money laundering is any activity aimed to hide the origin and/or the destination of a flow of money in order to reduce the probability of sanctions. In order to describe the economics of money laundering, the starting point is the definition of its microeconomic foundations, which are based on the existence of a rational actor who derives revenues from a criminal activity and from the assumption that his/her expected utility depends on four key elements: expected revenues, laundering costs, likelihood of being caught, and magnitude of the sanction.

The micro basis of the money laundering can explain its macroeconomic effects. Money laundering can function as a multiplier mechanism of the weight of the illegal sector in a given territory or country. In order to prevent and combat the polluting effects of money laundering, an effective regulation has to be designed, based on a correct incentives alignment between the supervisors and the financial intermediaries.

## Definition

*Money laundering is any activity aimed to hide the origin and/or the destination of a flow of money in order to reduce the probability of sanctions.*

Only recently the economic analysis has developed a peculiar focus on the financial issues related to the study of crime, thus far completely absent (Masciandaro 2007; Unger and Van der Linde 2013). In this entry, a simple framework to explain the micro foundations of the money laundering activities in order to analyze its macroeconomic effects on the relationship between the illegal activities and the economic sector as a whole will be offered.

## Microeconomics

The emphasis on the study of money laundering has progressively increased, recognizing its potential role in the development of any crime that generates revenues and or in financing a crime. In fact, the conduct of any illegal activity may be subject to a special category of transaction costs, linked to the fact that the use of the relative revenues increases the probability of discovery of the crime and therefore incrimination. The analysis zooms on the economics of concealing illegal sources of revenues, but the same reasoning can be applied in discussing the hidden financing of illegal activities (money dirtying).

However, we should stress that in terms of economic analysis, the financing of illegal activities (money dirtying) is a phenomenon not perfectly equivalent from the recycling of capital (money laundering).

The money dirtying resembles money laundering in some respects and differs from it in others. The objective of the activity is to channel funds of any origin to individuals or groups to enable illegal acts – for example, terrorism – or activities. Again in this case, an organization with such an objective must contend with potential transaction costs, since the financial flows may increase the probability that the financed crime will be discovered, thus leading to incrimination. Therefore, an effective money dirtying action, an activity of concealment designed to separate financial flows from their destination, can minimize the transaction costs.

The main difference between money laundering and money dirtying is in the origin of the financial flows. While in the money laundering process the concealment regards capitals derived from illegal activity, the illegal actor or organization can use both legal and illegal fund for financing their action.

Money laundering and money dirtying may coexist. A typical example is the financing of terrorism with the proceeds from the production of narcotics. In those specific situations, the importance of the transaction costs is greater, since the need to lower the probability of incrimination concerns both the crimes that generated the financial flows and the crimes for which they are intended. The value of a concealment operation is even more significant.

The definition of money laundering points up its specialness with respect to other illegal or criminal economic activities involving accumulation and/or reinvestment. Now, given that the conduct of any illegal activity may be subject to a special category of transaction costs, linked to the fact that the use of the relative revenues increases the probability of discovery of the crime and therefore incrimination, those transaction costs can be minimized through an effective laundering action, a means of concealment that separates financial flows from their origin.

In other words, whenever a given flow of purchasing power that is potential – since it cannot be used directly for consumption or investment as it is the result of illegal accumulation activity – is transformed into actual purchasing power, money laundering has occurred.

Focusing our attention on the concept of incrimination costs enables us to grasp not only the distinctive nature of this illegal economic activity but also its general features. The definition here adopted maintains basic unity among three aspects that, according to other points of view, represent three different objects of the anti-laundering action: the financial flows (layering), the wealth and goods intended as terminal moments of those flows (integration), and the principal actors, or those who have that wealth and goods at their disposal (placement).

In this general framework of analysis, there will always be an agent who, having committed a crime that has generated accumulation of illicit proceeds, moves the flows to be laundered, so as to subsequently increase his/her financial assets, by investment in the legal sector or re-accumulation in the illegal sector. The agent can be an individual or a criminal organization.

By criminal organization, we mean a group of individuals and instrumental assets associated for the purpose of exclusively exchanging or producing services and goods of an illicit nature or services and goods of a licit nature with illicit means or of illicit origin.

In general, following the classic intuition à la Becker, it can be claimed that the choices of an economic agent to invest his/her resources in illegal activities – as money laundering is – will depend, ceteris paribus, on two peculiar magnitudes, given the possible returns: the probability of being incriminated and the punishment he/she will undergo if found guilty.

Now, to undertake money-laundering activity, the agent possessing liquidity coming from illegal activity will decide whether to perform a further illicit act, in a given economic system – i.e., money laundering – assessing precisely the probability of detection and relative punishment and comparing that with the expected gains, net of the economic costs of this money-laundering activity.

The choice of the agent requires that the crime in question, and the relative production function, be basically autonomous with respect to other forms of crime, those that generated the revenues in the accumulation phase.

Assigning a monetary utility to the crime of money laundering, by giving it a unitary expression, actually summarizes the value of a series of more general services that stimulate the growth of demand for money-laundering services on the part of the agents that accumulate illegal resources. Money laundering, in fact, produces for its users:

1. The economic value, in the strict sense, of minimizing the expected incrimination costs, transforming into purchasing power the liquidity deriving from a wide range of criminal activities (transformation); transformation, in turn, produces two more utilities for the criminal agent.

2. The possibility of increasing his/her rate of penetration in the legal sectors of the economy through the successive phase of investment (pollution).
3. The possibility of increasing the degree to which the criminal actors and organizations are camouflaged in the system as a whole (camouflaging).

Having defined the micro choices in the most general terms possible, we can now investigate the macro effects of money-laundering choices.

## Macroeconomics

To define a macro model of the accumulation-laundering-investment process, we focus on the behavior of a general criminal sector that derives its income from a set of illegal activities and that, under certain conditions, must launder the income to invest it. We will highlight the role of money laundering as an overall multiplier of the criminal sector endowment.

Let us assume – as in Masciandaro 1999 and then in Barone and Masciandaro 2011 – that in a given economic system, there is a criminal sector that controls an initial volume of liquid funds *ACI*, fruit of illegal activities of accumulation. Let us further suppose that, at least for part of those funds, determined on the basis of the optimal microeconomic choices already discussed in the previous pages, there is a need for money laundering. Without separating these funds from their illicit origin, given the expected burden of punishment, they have less value. Money-laundering activity is therefore required.

To underscore the general nature of the analysis, we can claim that the demand for money-laundering services could be expressed – distinguishing the different potential components of a criminal sector according to their primary illegal activity – by organized crime in the strict sense, by white collar crime, or by political corruption crime, also considering the relative crossover and commingling.

Each laundering phase has a cost for the criminal sector, represented by the price of the money-laundering supply. The price of the money-laundering service, all other conditions being equal, will depend on the costs of the various money-laundering techniques. Let us suppose that in the money-laundering markets, the criminal sector is price taker and that the cost of money-laundering $cR$ is constantly proportional to the amount of the illicit funds; designating the costs with $c$, both regulatory and technical, we can write:

$$cR = cACI \tag{1}$$

If the first laundering phase is successful, the criminal sector may spend and invest the remaining liquid funds $(1 - c)yACI$ in both legal economic activities (investment) and illicit activities (re-accumulation).

The trend to use specialist money launderers – with their explicit or implicit fees – is increasing. These operators use their expertise to launder criminal proceeds. In general, the professionals may be witting or unwitting accomplice; but in any case, the buildup of the overall procedure represents a cost.

We assume that in general, the money-laundering procedures represent a cost for organized crime, notwithstanding it is well known that the criminal groups can implement legal businesses also for the concealment of their illegal proceeds and these businesses can produce profits. As it will be evident below, the smaller the money-laundering costs will be, the greater the multiplier effect.

The criminal sector will spend part of the laundered liquidity in consumer goods, equal to $d$, while a second portion will be invested in the legal sectors of the economy, for an amount of $f$, and then a third portion, equal to $q$, will be reinvested in illegal markets (giving, of course, $d + f + q = 1$).

On the one side share of illegal funds needs to be spent: minimizing incrimination risks comes at a price; the criminal sector has to pay a price. On the other side, we suppose that a share of dirty money will be reinvested in the illegal market without concealment. For example, in all illicit services, cash is by definition the currency of choice, running in a closed circuit separate from the legitimate market.

Reinvestment in criminal markets is a distinctive feature of actual organized crime groups, given their tendency toward specialization. Organized crime tends to acquire specialist functions to augment their illegal businesses.

If the criminal sector makes investment choices according to the classical principles of portfolio theory, indicating with $q(r, s)$ the amount of laundered funds reinvested in illegal activities, with $r$ the actual expected return on the illegal re-accumulation, and with $s$ the relative. Finally, we can assume that the re-accumulation of funds in the illegal sector requires their laundering only in part, thus indicating with the positive parameter $y$ the portion of illegal re-accumulation that requires laundered liquidity.

The criminal sector reinvests both clean and dirty money, and then a new flow of illegal liquidity will be created. The illegal revenues will be characterized again by incrimination costs, which will generate a new demand for money-laundering services. It will be therefore equal to:

$$(1+r)(1-c)^2 qy^2 ACI \tag{2}$$

The crucial assumption is that both the lawful investment and part of the unlawful re-accumulation require financing with "clean" cash. This assumption can be supported by the presence of rational, informed operators in the supply of services to the criminal sector for the illegal re-accumulation or by rationality of the criminal himself/herself, who wishes to minimize the probability of being discovered.

Repeating infinite times, the demand for money-laundering services, which each time encounter a parallel supply, with the values of the parameters introduced remaining constant, the total amount of financial flows generated by money-laundering activity $AFI$ will be equal to:

$$AFI = \frac{yACI\,(1-c)}{1-yq\,(1-c)\,(1+r)} = mACI \tag{3}$$

with $0 < c, q, y < 1$.

The flow $AFI$ represents the overall financial endowment generated by the money-laundering activity, and $m$ can be defined as the multiplier of the model. Doing comparative static exercises, it is easy to show that the amount of liquidity laundered will increase as the price of the money-laundering service declines:

$$\frac{\partial AFI}{\partial c} = -\frac{ACIy}{[1-qy(1-c)\,(1+r)]^2} < 0 \tag{4}$$

Therefore, the more effective the money-laundering action, the greater the cash flows available to the criminal sector for reinvestment, illegal and legal, will be. Money laundering represents the multiplier of the illegal sector.

## Regulation

Summing up the key features of money laundering, the micro foundations have been based on the existence of a rational criminal actor or organization, which derives revenues from an illicit or criminal activity. The criminal actor or organization wants to maximize the expected utility of his/her or its illicit proceeds. The expected utility increases with the average return, and it decreases with the costs for laundering, with the probability of being caught, and with the severity of the sanction when being caught.

Further, the macroeconomic effects of money laundering have been captured using a multiplier model. Money laundering triggers a multiplier process, which ends up with higher laundering and higher criminal activities. Money laundering harms the economy. Therefore, it can be useful to design and implement a regulation to prevent and combat the money-laundering phenomena.

On this respect, the starting point has to be to recognize that the banking and financial industry usually play a pivotal role for the development of the criminal sector as a preferential vehicle for money laundering.

The main actors are on the one hand the regulatory agents, who want to combat money laundering, and on the other hand the financial intermediaries, who can be either honest and compliant or dishonest and noncompliant. Asymmetric information and principal-agent problems are typical for this market. The design of anti-money-laundering regulations must take four aspects into consideration: the difference in information assets between the individual intermediaries and the agency, the non-verifiability of bankers' efforts to comply, the costliness of that effort for the intermediaries, and the non-verifiability of the influence of the effort on the performance of the regulation.

To deal with such difficulties, the best way of analyzing the regulatory issues is to implement a principal-agent methodology, managing the incentive problems that arise at least in a three-layer hierarchy, which includes public authorities, financial institutions, and supervisors. It is possible to show – Dalla Pellegrina and Masciandaro (2009) – that the under asymmetric information, the effectiveness of the regulation depends on three crucial conditions.

First, the participation constraint of the financial institutions requires that the incentive scheme is well balanced, meaning that both rewards and penalties must be defined, in order to minimize the difference between the private costs in implementing a regulatory model and the public gains in collecting useful information against money laundering.

Second, excessive fines per se do not necessarily provide incentives to the financial institutions to improve their action. In particular, given the incentive scheme of the financial institutions, the quality of the supervision can be a good substitute for the severity of punishments: the more effective the (potential) supervisory action in monitoring ex post the money-laundering risk, the more likely the effectiveness of the financial institutions in building up ex ante their monitoring models.

Third, other things being equal, if the cost of supervision depends on its quality, also the efficiency of the supervisory agencies matters. Again, the importance of the quality and the efficiency of supervision can be particularly relevant.

## References

Barone R, Masciandaro D (2011) Organized crime, money laundering and legal economy: theory and simulations. Eur J Law Econ 32(1):115–142

Dalla Pellegrina L, Masciandaro D (2009) The risk based approach in the new European anti-money laundering legislation: a law and economics view. Rev Law Econ 5(2):932–952

Masciandaro D (1999) Money laundering: the economics of regulation. Eur J Law Econ 7(3):225–240

Masciandaro D, Takats E, Unger B (2007) Black finance. The economics of money laundering. Edward Elgar, Cheltenham

Unger B, Van der Linde D (eds) (2013) Research handbook of money laundering. Edward Elgar, Cheltenham

# Multiple Tortfeasors

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

## Definition

Multiple tortfeasor issues cover cases where the loss is jointly caused by several people acting in a common purpose or not. A lot of examples illustrate these situations including accident law, environmental damage, product liability, etc. The law and economics literature has devoted much attention on these situations in order to understand the properties of different liability rules. Two levels of discussion should be distinguished: first, the negligence/strict liability debate, and second, the joint or no-joint liability rules meaning that the victim may get compensation back by any of the tortfeasors. The article surveys the most important results in law and economics and insists on the fact that multiple tortfeasor cases lead to paradoxes which challenge the very basis of the tort law and the economic rationality.

### From One Defender to Several

At the very beginning, tort law has been one of the most promising fields in law and economics: Coase, Calabresi, and Posner were particularly interested in the efficiency of liability rules.

However, the economic basic model implied only one tortfeasor (Posner 1973; Coase 1960; Calabresi 1970). The scope of this model was limited and did not cover more complex cases where several tortfeasors jointly cause the harm suffered by the victim, what is called multiple tortfeasor issues. Sometimes, cases of multiple tortfeasors refer to harm due to a common purpose of several people. However, the scope of multiple tortfeasors is broader and covers also harm due to independent wrongdoers' behaviors. Many examples of multiple tortfeasors come in mind: accident law, environmental law, products liability, etc. the law and economics literature about multiple tortfeasor cases started at the beginning of the 1980s and does not necessarily converge on clear principles. The reason is that two levels of discussion are involved: the first is about the regime of liability (strict liability or negligence), the second about the joint or non-joint liability.

To handle with this complexity, legal scholars have fallen into the habit of forming their reasoning based on typologies to classify different cases (Hart and Honoré 1985; Landes and Posner 1980, 1983). The first case is simple: several tortfeasors independently act at the same time, and their joint behaviors simultaneously lead to harm (two polluters pouring toxic waste in a river). The second case is sequential: a first tortfeasor causes at time $t_1$ a first harm that is enhanced at time $t_2$ by another wrongdoer, *etc*. The third case covers the example of contribution from the victim which has participated to its own harm (e.g., because he has not taken sufficiently care level). The fourth case implies uncertain causation: harm occurred without knowing with certainty who is the true wrongdoer (two hunters shooting while only one bullet actually harms the victim). In the following, we focus on the third first cases, and we do not deal with uncertain causation (see the article ► "Causation" in the *Encyclopedia*).

These complex cases have been extensively studied by legal scholars (Hart and Honoré 1985; Wright 1985; Stapleton 2013) because they challenge the usual way to consider liability, causation, and torts. Indeed, imagine two people lighting a fire in a forest. Suppose that these two fires merge together and destroy the house of the victim. In such a case of overdetermined causation, the "but for test" criterion advocated by the law is useless: each of the tortfeasors could escape from his responsibility by pretending that, without his own action, the harm would have occurred anyway. It would be unfair and inefficient to follow the black letter of the "but for test," and these cases require other legal solutions.

From the economic point of view, these cases are also difficult to handle with. We will survey one of the most key elements of the findings of law and economics literature on this topic. First, we insist on the efficiency and incentives aspect. We also deal with the fairness of the sharing rules applied by judges to apportion damage. Then, we compare three rules from an economic perspective: several liability, joint and several liability, and channeling liability. Last, we conclude with dealing with some of the most striking paradoxes raised by multiple tortfeasor cases.

## Incentives, Efficiency, and Fairness

As we will see, in many instances, multiple tortfeasor cases challenge the result of efficiency which holds for a simple one victim/one tortfeasor case. First, it is now well-known that one of the most important findings in law and economics is lost when several tortfeasors are implied. In a paper published in 1981 by Aivazian and Callen, and entitled "Coase Theorem and the Empty Core," these authors take a simple example where two polluters cause harm to a victim. Under the general conditions of the Coase theorem – perfect delineation of property rights and zero transaction costs – the result of invariance and efficiency cannot be proved. The reason is that the core of the game may be empty meaning that there is no stable allocation which could emerge from negotiation: victim and tortfeasors negotiate again and again without any convergence of their offers.

In case of positive transaction costs, liability rules are studied to implement an efficient output in terms of care, activity, and actual harm (Shavell 2004). But when it comes to efficiency in cases of multiple tortfeasors, it is difficult to conciliate two principles: the principle that every tortfeasor should pay for the harm he is responsible for and the fact that the total amount of compensatory damage be equal to the harm. Most of times, if principle 2 is followed, underdeterrence is expected. On the contrary, if principle 1 is followed, it requires implementing a sanction which is above the amount of harm. This result is easy to show in case of strict liability. Assume that two tortfeasors 1 and 2 cause harm to a victim. The objective function of social cost to be minimized is $x_1 + x_2 + p(x_1,x_2)$.h, with $x_1$ the cost of care of 1, $x_2$ the cost of care of 2, $h$ the amount of harm, and $p$ the probability of harm. To induce optimal care, it would be necessary that each tortfeasor bears damage of $h$. This is impossible insofar as $h$ will be split among the two liable tortfeasors (usually, the compensatory damage is equal to harm, no more, no less). Other authors have elaborated on this issue by providing alternative rules. For example, Miceli and Segerson imagine a rule where each tortfeasor is responsible for the marginal damage that it causes. In that case, efficiency is achievable, but it requires that the total amount jointly paid by tortfeasors be above the loss caused. Miceli and Segerson consider that a mixed system requiring on the one hand, the recovery of the loss to the victim and, on the other hand, a fine to be paid to the state could achieve efficient incentives (Miceli and Segerson 1991; Miceli 1997).

While under strict liability, tortfeasors are threatened to pay less damage than required by efficiency; under negligence rule, incentives are different. If the efficient care standard ($x_1$*, $x_2$*) is implemented by the law, a tortfeasor will not be held responsible if his level of care is above the efficient level. The best way to analyze the behaviors of tortfeasors is to use game theory and to wonder whether the equilibrium of the game exists and reaches social efficiency. If tortfeasor 1 has chosen $x_1$*, the best response of tortfeasor 2 is to choose the efficient level $x_2$*. If not, he would pay for the entire damage. The situation is symmetric fort the agent 1. As Shavell states, “The superiority of the negligence rule has to do with the fact that, by the nature of the negligence rule, each party is threatened with damages equal to the *entire* harm $h$ when other parties act optimally; whereas under strict liability, each party is threatened with a lesser amount” (Shavell 2007). This result holds when tortfeasors are jointly liable (they expect to pay for the entire loss) but does not hold in case of several liability where they would expect to bear only a share of the loss (Kornhauser and Revesz 1989). In the previous paragraphs, we have implicitly supposed that all the tortfeasors are solvable. In case of insolvency, the previous results may be altered, and we will focus on this issue in part III.

In a different perspective, it is possible to have a more fairness-oriented point of view and be interested in the contribution of tortfeasors. The idea is to have a scheme of payment which is in line with the causal contribution of each tortfeasor. Some legal scholars have elaborated on this idea of causal contribution (Stapleton 2013). In law and economics, comparative causation is such a rule (Parisi and Singh 2010). This idea dates back to Calabresi for whom negligence has an unexpected effect to promote a one or no liability. He suggests splitting the loss among parties following their causal contribution. Parisi and Singh follow this argument and have carefully studied the properties of a causal contribution rule mixed with a negligence rule. They show that such a rule has interesting properties in terms of optimal care level and optimal activity level and consider that it is a fair and equitable rule. In a different perspective focused on ex post issues, other scholars use cooperative game theory and consider that the Shapley value is a good candidate to provide an evaluation of the contribution of each tortfeasor involved in the case of indivisible harm. Moreover, the Shapley value has axiomatic properties interesting for the law, and it could be shown that most of principles advocated in the *Restatement of Tort* (American Law Institute 2012) in the USA to apportion damage are close to the elementary principles of the Shapley value (Ferey and Dehez 2016a).

### Several, Joint and Several, and Channeling Liabilities

Even though it was possible to clearly apportion harm among tortfeasors, a separate issue raises about how the victim could trigger responsibility against each of them. In most countries, three rules are implemented by the law: several liability, joint and several liability, and channeling liability (Faure 2016). Several liability states that each tortfeasor is responsible for his share and the victim has a separate claim against each of them but only for their respective shares. Several liability has two consequences: first, the victim has to bear the costs of several suits to recover all the compensation of his harm from all the tortfeasors separately; second, the risk of insolvency of one of the tortfeasors is bear by the victim.

Several and joint liability is different: the victim has a claim for the entire compensatory amount against any of the tortfeasors and gets all its recovery damages back from him. Then, in most of legal systems, the tortfeasor who has compensated the victim has a claim against the other tortfeasors to get their respective shares of responsibility back. Joint and several liability has two consequences: first, litigation costs are decreasing for the victim (and increasing for the defender); second, it is expected that the victim will choose the most wealthy defender to get its recovery back (following a deep pocket argument), and the risk of insolvency is bear by the solvable defender, not by the victim. A lot of fields in torts around the world are organized under a joint and several liability. In Europe, for example, the *European Principles of Tort Law* (see European Group on Tort Law 2005) mainly advocate joint and several liability.

Third, channeling liability indicates ex ante who among the potential tortfeasors is fully responsible and most of times without any claim against the others. Channeling liability is implemented by some international conventions. For example, the *Paris Convention on Nuclear Liability* has decided that nuclear liability is channeled to operators. The relative properties of the three rules have been extensively discussed in the law and economics literature which focus on four main topics: litigation costs, incentives, insolvency, and insurance (Faure 2016). We could add that the choice of one of the three rules (several, joint and several, or channeling liability) may be influenced by pressure groups without any global efficiency concerns.

Regarding litigation costs, the three rules are different. Channeling liability makes recovery easier for the victim: the identity of the party to be sued is certain, and only one lawsuit has to be filed. On the contrary, several liability seems costly for the victim who is required to pay for as much suits as the number of tortfeasors. Joint and several liability is intermediate: the litigation costs for the victim to get compensation back is relatively low but increase for the deep pocket tortfeasor which is required to pay for lawsuits against the other tortfeasors. And it could be expected that enforcement of the law will be more effective if litigation costs are low for the victim.

Regarding insolvency, the three rules have also different properties. The insolvency issue arises when at least one tortfeasor is not able to pay his share of damage back. Full solvency of tortfeasors seems to be a strong hypothesis which does not hold in many cases, and insolvency should be expected due to the high amounts that firms have to pay. First is the distributive issue: who should bear the risk of insolvency, the victim or the other tortfeasors? Most of legal systems consider that it is fair enough to avoid the undercompensation of the victim due to insolvency. The risk of non-recovery is shifted to the solvable tortfeasor.

This point has consequences on incentives. Under joint and several liability with insolvent actors and channeling liability, a tortfeasor may be asked to pay beyond his own share without any possibility of recovery from other defenders. The consequences on the level of care and the level of activity taken by the parties are not clear. First, the insolvent tortfeasor does not take into account the risks he created above the total value of its assets. Up to a certain point, he is underdeterred. Second, the solvent tortfeasor may be aware of this insolvency and internalizes the fact that, in case of several and joint liability, he should pay for the other's share. The main effect is on the activity level which may be reduced or eliminated to avoid any liability. This effect is called the crushing effect of liability. Third, the opposite effect may also happen. Intuitively, the reasoning insists on the fact that, in case of limited solvency, liability rule has no deterrent effect above a certain level anymore (it has the same effect as statutory caps on liability). Up to a certain point, a domino effect occurs: the share of the insolvent tortfeasor to be paid by the remaining solvable actor is so huge that liability has no deterrent effect anymore on the solvent agent (Kornhauser and Revesz 1989). Lastly, a monitoring policy may be expected from parties: a tortfeasor who knows that he is likely to pay for others has an incentive to control and monitor other parties to be sure that they decide optimal levels of care and/or activity. However, in some cases, a defender has no way to monitor the level of care taken by other.

The three rules have strong consequences on insurance markets. The fact that one tortfeasor pays more than his share means that its insurance company is not able to calculate a risk premium anymore. Virtually, the insurance company will guarantee all the losses caused by all the participants of the market. Here is the risk of a crushing effect on the insurance market. The argument could be mitigated by the diversification of risks: It is unlikely that the same insurance company be concerned by all the cases implying multiple tortfeasors (Faure 2016).

### Paradoxes

We would like to conclude by several paradoxes raised by multiple tortfeasor cases which illustrate significant challenges: overdetermined causation, aggregation issues, and offsetting benefits.

A first series of paradoxes is well-known by legal scholars: the overdetermined cases (Hart and Honoré 1985). It is the case of the two fires merging and destroying a property. In that cases, the black letters of the classical criterion of causation leads to irrational results. That is why other views on causation are proposed by legal scholars like the NESS test – necessary element of a sufficient test – (Wright 1985) or the causal contributions (Stapleton 2013). These criteria of causation may be modeled in an economic framework (Ferey and Dehez 2016b) and explain the reasons of the paradoxes.

Implying several people, multiple tortfeasor cases lead also to aggregation puzzles. In a recent paper, Porat and Posner have provided a general framework to capture aggregation issues. Regarding torts, they insist on the general consequences of the burden of proof. As many scholars, they consider that preponderance of evidences rule may be translated in terms of a probability superior to 50%. Imagine that a victim suffers two harms: the first one is due to an error of the surgeon of an hospital with a probability of 0,6 (event A) and a second one by a nurse of the same hospital (event B) with a probability of 0,6 (Porat and Posner 2012). Then, the victim files a lawsuit against the hospital for the two harms. Without aggregation, liability of the hospital will be held for the two harms because the probability of event A and of event B is superior to 0,5. However, it could be said that the probability that the hospital be responsible jointly for the two harms is only of 0,36 (0,6*0,6) which is the probability of the conjunction of the two events. The paradox raises from the fact that the preponderance of evidence leads to consider that the hospital should not be held responsible for the two harms. Such paradoxes are particularly interesting regarding multiple tortfeasors. A general typology of all the different types of aggregation issues is provided in Porat and Posner (2012).

The third type of paradox is implied by what is called “offsetting benefit”. Offsetting benefit issues are broader than multiple tortfeasor issues, but they are particularly concerned with. Porat and Posner have extensively dealt with these cases (Porat and Posner 2014). A simple example of the problem is the following. Imagine a victim driving his car to reach the airport where he has to take a flight. On the road to the airport, he is harmed by a wrongdoer and cannot take his plane. It then appears that the plane crashes. Following strict causation reasoning, it could be said that the wrongdoer has caused an injury but has also saved the life of the victim. Should we consider that the benefits due to the accident has to be taken account in the calculus of the compensatory damage? In this example, the victim could even compensate the tortfeasor due to the behavior of a potential second tortfeasor. There are different ways to solve these types of cases, and Porat and Posner provide some elementary principles to guide courts.

To conclude, it is clear that a lot needs to be done to better understand all the aspects of multiple tortfeasors issues. We have shown how the difficulties of this topic are both philosophical, legal, and technical, implying different aspects in terms of efficiency and fairness and let a room open for political choices.

## Cross-References

- ► Causation
- ► Strict Liability Versus Negligence

## References

Aivazian VA, Callen JL (1981) The Coase theorem and the empty core. J Law Econ 24:175–181

American Law Institute (2012) Restatement (Third) of the law of torts: liability for physical and emotional harm. Executive Office, American Law Institute, Saint Paul

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

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

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

Faure M (2016) Attribution of liability: an economic analysis of various cases. Chic-Kent Law Rev 91:603–636

Ferey S, Dehez P (2016a) Multiple causation, apportionment and the Shapley value. J Leg Stud 45:143–171

Ferey S, Dehez P (2016b) Overdetermined Causation, Contribution and the Shapley Value. Chic-Kent Law Rev 91:637–658

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

Kornhauser LA, Revesz RL (1989) Sharing damages among several tortfeasors. Yale Law J 98:831–884

Landes WM, Posner RA (1980) Joint and multiple tortfeasors: an economic analysis. J Leg Stud 9: 517–555

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

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

Miceli TJ, Segerson K (1991) Joint liability in torts: marginal and infra-marginal efficiency. Int Rev Law Econ 11:235–249

Parisi F, Singh R (2010) The efficiency of comparative causation. Rev Law Econ 6:219–245

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

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

Posner RA (1973) Economic analysis of law. Little Brown, Boston

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

Shavell S (2007) Liability for accidents. In: Shavell S, Polinsky M (eds) Handbook of law and economics, vol 1. Elsevier, Oxford, pp 142–182

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

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

# Music Royalty Rates for Different Business Models: Lindahl Pricing and Nash Bargaining

Marcel Boyer[1,2] and Anne Catherine Faye[3]
[1]Department of Economics, Université de Montréal, Montréal, QC, Canada
[2]Toulouse School of Economics, Toulouse, France
[3]Analysis Group, Montréal, Canada

## Definition

The Lindahl Equilibrium and Nash Bargaining Solution serve as useful and complementary analytical tools that provide guiding principles for the determination of appropriate music royalty rates in the current digital era, which poses challenges and pitfalls for the pricing of information goods, most notably musical works and sound recordings.

## Introduction

We discuss the economic concepts of Lindahl Equilibrium and Nash Bargaining Solution as useful and complementary analytical tools in the context of the current digital era, which poses challenges and pitfalls for the pricing of information goods, most notably musical works and sound recordings. These concepts of modern economic theory are more than ever useful, even required, to provide the guiding principles underlying the determination of appropriate music royalty rates.

There is a general agreement between rightsholders, copyright users, and tariff-setting organizations that tariffs should be "fair and equitable" to both rightsholders and users and reflect the value or benefits the users derive from copyrighted works.

From an economic perspective, a "fair and equitable" level of transactions and compensation is equivalent to the level of transactions and compensation that emerges from a competitive market where willing buyers and willing sellers freely negotiate and settle transactions. Such willing buyers and sellers are assumed to be price-takers or devoid of market power. They would agree on transactions up to the point where the marginal value of an additional transaction for buyers is equal to the marginal cost of that additional transaction for sellers. From that perspective, the tariff to be paid for the use of copyrighted musical works and sound recordings should be based on the amount that users would willingly pay if they were transacting in a well-functioning competitive market.

The Copyright Board of Canada, the US Copyright Royalty Board, and other tariff-setting institutions represent a surrogate for a competitive market where the price that would prevail for copyrighted works is determined, if such a market existed and operated efficiently. To determine a competitive price for copyrighted works, these institutional bodies typically consider relevant proxies or indicators on buyers, sellers, prices of substitute products and services, industry characteristics, and virtual or simulated competitive processes such as auctions, etc. The challenges encountered in emulating a competitive market are numerous, some of the most salient being the fact that musical works and sound recordings are information goods and that digital technologies are profoundly changing the copyright landscape.

## Information Goods and Digitization: Lindahl Pricing and Nash Bargaining

From an economic perspective, musical works and sound recordings are "information goods." An information good is a good whose consumption by one consumer does not prevent its consumption by others, a characteristic of "information" under different forms. In other words, once a musical work or sound recording is created, it can be "consumed" by all, as one person's use of a musical work or sound recording does not prevent its simultaneous or subsequent use or consumption by someone else. See for example, Varian (1999) and Bakos et al. (1999).

Although related, the "information good" nature of musical works and sound recordings and the digitization of music are two different but equally challenging factors in the current copyright setting. The first one relates to the indefinite sustainability of a product, as one's consumption does not destroy the consumed unit, which remains fully and unabatedly available for everyone else now and in the future, hence making musical works and sound recordings akin to non-decaying assets. The second one relates to the much-reduced dissemination cost of that product or asset in a digital context. See Boyer (2018).

In such a context, striking the right balance between creators' rights and users' interests is a difficult and multifaceted endeavor for different reasons: (a) musical as well as other cultural products such as literary works are costly to create and (b) digital technologies have significantly reduced the marginal cost of dissemination of those copyrighted works.

Due to these factors, pricing copyrighted musical works and sound recordings, with the objective of achieving proper compensation for creators, i.e., a creators' right, *and* maximal dissemination of such goods, i.e., a users' right, requires a move away from the usual analysis aimed at setting a product's price equal to its marginal cost. Setting a price equal to marginal cost would clearly not enable the proper, fair, or competitive compensation of sellers, producers, and creators. Considering that musical works and sound recordings are permanent assets rather than perishable consumption goods raises questions regarding the proper price concepts to use: the same unit can be sold and resold infinitely many times.

The determination of relevant copyright tariffs rests not so much on the cost of creation, which underlies the supply function of new musical works and sound recordings, but rather on the value of such goods for the users. Indeed, the Copyright Board of Canada (2002) acknowledged that "the important notion in information industries [is that] pricing tends to be based on the value to the buyer, not on cost to produce."

The supply function of musical works and sound recordings in the rightsholders' collective repertoire (stock) is horizontally constant at a price $p$ just above 0 (infinite price elasticity at $p = \varepsilon$). That is because the marginal cost, which underlies the supply function, is quasi-zero for the rightsholders: zero marginal cost of reproduction and dissemination and zero marginal opportunity cost, those musical works and sound recordings being resalable an indefinite number of times. Setting the tariff equal to this marginal cost would generate no revenue for the seller, here the rightsholders, thereby failing to meet a central objective of the copyright institution. Insofar as the compensation of rightsholders must come through transactions, not from public subsidies or grants, the setting of tariffs is cast in a second best economic framework.

In this context, the proper price equilibrium is the Lindahl (1958) equilibrium. It is a generalization of the competitive equilibrium for public or information goods, according to which, given a stock of musical works and sound recordings, different users pay for access to a relevant repertoire, prices equal, or proportional to their respective derived marginal values.

The notion of Lindahl pricing was developed to characterize both the optimal or efficient quantity of a public good and a way to finance its production cost. It requires that the price paid by a given buyer or user of a public (or information) good be positively linked to the value or the amount of satisfaction derived from the consumption of that public good, as everyone is assumed to consume, at least virtually, the whole good. The user deriving a greater (marginal) value from the same marginal unit would pay a higher price. As long as the sum of those prices is above the marginal cost, additional units should be produced. The efficient quantity and quality level are reached when the sum of users' marginal values (prices) is equal to the marginal cost. The Supreme Court of Canada has asked the Copyright Board of Canada to consider such analytic framework in setting tariffs.

## Technological Neutrality

The Supreme Court of Canada, in its 2015 landmark decision introducing the principles of technological neutrality and balance, stated:

> One element of just compensation is an appropriate share of the benefit that the user obtains by using reproductions of their copyright-protected work in the operation of the user's technology. That just compensation must be valued, however, in accordance with the principle of technological neutrality. While highly unlikely, where users are deriving the same value from the use of reproductions of copyright-protected works using different technologies, technological neutrality implies that it would be improper to impose higher copyright-licensing costs on the user of one technology than would be imposed on the user of a different technology. To do so would privilege the interests of the rights holder to a greater degree in one technology over the other where there is no difference between the two in terms of the value each user derives from the reproductions.
>
> The converse is also true. Where the user of one technology derives greater value from the use of reproductions of copyright-protected work than another user using reproductions of the copyright-protected work in a different technology, technological neutrality will imply that the copyright holder should be entitled to a larger royalty from the user who obtains such greater value. Simply put, it would not be technologically neutral to treat these two technologies as if they were deriving the same value from the reproductions. (Supreme Court of Canada 2015, paragraphs 70, 71)

The Supreme Court asked the Copyright Board to consider that if a user with a given technology derives more value from a product, say a repertoire of musical works and/or sound recordings assets, than another user with a different technology, the former user should pay more than the latter, even if both use the same repertoire. In other words, the two users would access the same asset at different prices.

The Supreme Court thus recognized that, for efficiency and fairness reasons, users with different business models may be subject to different tariffs for access to the same stock of musical works and sound recordings. There is a direct link between technological neutrality and Lindahl principles for pricing information goods.

Applying Lindahl principles is arduous and full of potential pitfalls especially in the context of copyright royalties. First, determining the (marginal) values different users assign to an asset is challenging. Second, many tariffs are expressed as a proportion or percentage of an accounting base rather than as a price per unit.

Finding the (marginal) values different users attach to or are willing to pay for an asset could be achieved through simulated auctions. The Copyright Board has alluded to such auctions in its 2002 Digital Pay Audio Decision as follows: "The Board finds it useful to comment on the 'simulated auction' approach which was discussed at the hearing. This scenario calls for setting the price of music at what one would be willing to pay to acquire a monopoly over DPA. That approach must be set aside because it focuses again on profitability at the expense of all else. This being said, the exercise is not without merit, if only because it highlights the important notion that in information industries, pricing tends to be based on the value to the buyer, not on cost to produce" (p. 8).

Many tariffs are expressed as a proportion or percentage of an accounting base due to the challenging factors involved in pricing copyrighted works on a per unit basis. Expressing royalties as a percentage of an accounting base, such as the user's revenues, serves different purposes. It allows for the following: (a) savings in transaction costs, including assessing and verifying copyright payments; (b) immunity to accounting manipulations if the rate base is well chosen and valid; and (c) risk-sharing between rightsholders and users as a whole because it may be difficult or impossible to know ex ante which user will be successful in turning profitable access to a stock of musical works and sound recordings. However, expressing royalty payments as a percentage of an accounting base does not provide any indication about the price of copyrighted work. This follows because a percentage is not a price.

## A Percentage Is Not a Price

A major source of confusion in royalty rate setting is the trap of "considering a percentage as a price." There is no reason to believe that the proper price to be paid for the same inputs, namely, the right to access a stock of musical works and sound recordings, would correspond to the same percentage of revenues irrespective of the characteristics of the underlying industries under consideration. In fact, using the same percentage will usually lead to subsidizing one industry at the expense of another because, under the economic law of one price, the same price for the same inputs used in two different industries will, in general, represent quite different percentages of the value of the industry outputs (or revenues). Moreover, Lindahl equilibrium calls for different prices for different users or consumers of the same good.

The following example can help illustrate why a percentage is not a price. Suppose an apartment in a low income housing project is valued at $100,000. Suppose also an apartment of the same size and configuration in a high-income housing project is valued at $1,000,000, i.e., ten times more. Then suppose that the same quantity and quality of paint is used for both apartments. As expected, the cost of painting the two apartments would be the same as the law of one price applies to the input market (i.e., paint). In economics, the law of one price states that similar products in the same market should sell at similar prices. Let us suppose that the actual cost of paint for the high value apartment is $1,000, i.e., 0.1% of the value of the apartment. If one takes the 0.1% as the "price" of the paint to be paid for the low-value apartment, one would get a price in dollar terms of 0.1% × $100,000 = $100 or one-tenth of the cost of the same amount and quality of paint used for painting the high-value apartment. In fact, the cost of paint expressed as a percentage of the value of the two apartments would be quite different: 1.0% for the low-value apartment and 0.1% for the high-value apartment, a difference by a factor of ten for the same quantity and quality of paint.

As we further discuss below, regulatory bodies responsible for setting royalties on the use of copyrighted works tend to avoid transferring percentages from one industry to another without proper adjustments. In doing so, they aim to ensure that percentages reflect similar prices for the same rights.

Making those different percentages correspond to Lindahl prices and percentages requires additional adjustments.

## Balance Between Rightsholders and Users

The Supreme Court of Canada has stated that achieving balance between the rights of creators and users requires the Copyright Board take both rights into account in the determination of royalty rates by considering "respective contributions of, on the one hand, the risks taken by the user and the investment made by the user, and on the other hand, the reproductions of the copyright-protected works to the value enjoyed by the user" (Supreme Court of Canada 2015, paragraph 75). The principle of balance is related to the economic concept of a bargaining game, which in turn is related to the concepts of competitive equilibrium and negotiated price.

A bargaining game is an analytical framework of game theory that seeks to model a situation in which there is a conflict of interest between different agents who have the opportunity to reach a mutually beneficial agreement but may nevertheless veto any agreement. When "there is more than one course of action more desirable than disagreement for all individuals but conflicting views over which course of action to pursue, then negotiations to resolve the conflict will take place" (Osborne and Rubinstein 1994, p. 117).

In this context, the framework of a bargaining game appears as a tool to model the negotiation process and to determine how different agents who contribute to the creation of a given surplus or value added can share such value among them. A solution to the bargaining game is a sharing formula that specifies what percentage of the total value added each agent receives at the end of the bargaining game. The total value to be shared, the agents' outside options (i.e., alternative options in case they do not reach an agreement, including the no-investment option), as well as their bargaining power all play a role in the determination of the solution.

John F. Nash Jr., the 1994 laureate of the Nobel Memorial Prize in Economic Sciences, proposed a solution to such a bargaining problem, known as the Nash Bargaining Solution (NBS). The NBS is derived from four axioms defined as desired or reasonable properties. Namely, the solution should be (Pareto) efficient, symmetric, immune to equivalent reformulations of players' objectives, and immune to irrelevant alternatives. The efficiency property simply states that the solution should leave no money on the table. The symmetry property states that if two agents are in similar positions or have similar capacities to negotiate, they should be treated equally, that is, obtain "similar" shares of the pie. Given that each party is rational, well advised, and controls an essential input, each holds similar power to negotiate and veto any solution and therefore can be considered to be equally capable of negotiating and affecting the solution. The other two axioms are more technical and not developed here.

Nash shows that there is only one solution to the above bargaining problem (Osborne and Rubinstein 1994, p. 307). In other words, there is a unique sharing formula that satisfies the four axioms: once each agent is properly compensated for the cost it incurs to sit at the negotiation table, the residual monetary value would or should be shared 50-50 between the agents. Given the reasonableness of its axioms, the Nash bargaining solution (NBS) is a powerful result. It says that whatever the negotiations conduct and/or process, i.e., no matter how the negotiation is conducted, the expected ultimate end-point or result can only be the NBS. As mentioned above, the NBS is closely related to the concepts of competitive equilibrium (willing buyer, willing seller) and negotiated price, two concepts regularly referred to in hearings before tariff-setting organizations.

The balance required by the Canadian Supreme Court to be considered in rate-setting must be reached in a context where standard perfectly competitive conditions do not prevail. Regarding copyrighted musical works and sound recordings, a standard competitive market with individual buyers and sellers with no market power and symmetric information does not generally exist. However, implicit negotiations can take place between representatives of the parties before a rate-setting body emulating a competitive framework and solution. A negotiated price would then be considered the solution of the bargaining game involving the different parties. Such negotiated prices are grounded in economics as accounting for (and therefore balancing) the relative contributions and alternate options of the parties involved in the negotiations. The Nash Bargaining Solutions of negotiations between different users and the rightsholders call, under the 50-50 rule, for different royalty payments for different users with different business models for packaging and transmitting or distributing music.

Hence, the principles of technological neutrality (Lindahl pricing) and balance of rights (Nash Bargaining) in the context of information goods require that different business models be charged different royalty rates for the same access to musical works and sound recordings. In other words, efficiency and fairness in royalty setting run against business model neutrality.

## The Case of Digital Pay Audio and Commercial Radio in Canada

In Canada, payments for communication rights of musical works and sound recordings in the digital pay audio industry (DPA) as well as in the commercial radio industry (CR) are measured as percentages of a rate base (the total revenue of the user). In fact, the Copyright Board (CB) has at numerous times compared and used as proxies royalty rates expressed as percentages of revenues by carefully applying relevant adjustments to account for differences among the industries considered.

Applying the above reasoning to the CR and DPA industries, similar prices for the same access rights to the same repertoire of musical works and sound recordings, when expressed in percentage terms, will represent a higher percentage in the lower value added DPA industry, where the main input is music, than in the higher value added CR industry, where musical works and sound recordings are one input among many including news, weather, or traffic reports and on-air personalities. In other words, similar prices would yield different percentages.

In its decision, the Supreme Court of Canada instructs the Board to consider that if a user, such as DPA services, derives more value from the product, say the repertoire of SOCAN Collective of authors and composers (musical works) and Re:Sound Collective of performers and makers (sound recordings) than another user, such as a CR operator, the former user should be asked “to pay more” than the latter.

In Canada, the commercial radio royalty rates for musical works and sound recordings are both 4.2% (before adjustments for repertoire) of the user’s revenues. Suggesting to increase the combined CR royalty rate of 8.4% to 10.6% for DPA on the basis that the latter uses 25% more music per given period, would clearly not satisfy the technological neutrality principle insofar as the DPA industry derives greater value from musical works and sound recordings relative to CR, which generates revenues from other programs.

One possibility would be to, first, derive the level of royalties generated by the 8.4% rule and, second, express it as a percentage of revenues generated by music only; indeed, the commercial radio industry depends less on music than the digital pay audio services industry. On average, music format radio stations broadcast music content during 80% of programming time and other content (news, survival programming, on-air personalities, or talk) during the remaining 20%, while DPA services play music 100% of the time, that is, 25% more. However, the CB has repeatedly expressed the view that on-air talent is more important than music to radio stations. In its 2002 DPA decision, it stated: “Radio may be designed around the use of music and musical genres but as a cost, and (probably) as a drawing card, on-air talent is far more important” (Copyright Board of Canada 2002, p. 8). If so, one may estimate that on-air talent generates far more and music far less than 50% of revenues, that is, say 60% and 40%.

As such, the resulting percentage rate of music royalties expressed on the basis of music-generated revenue would be 21.0 (= 8.4/0.4)%. Increasing that percentage rate by 25% to account for the larger use of music in DPA would yield an equivalent rate for DPA, i.e., 26.3%.

Hence, the percentages, which would satisfy both the principles of technological neutrality and balance, assuming that the commercial radio rates are properly set at their competitive market value level, would be, 8.4% of revenues for Commercial Radio and 26.3% of revenues for Digital Pay Audio services, before any adjustments for repertoire or other reasons.

## Cross-References

▶ Copyright
▶ Economic Efficiency

## References

Bakos Y, Brynjolfsson E, Lichtman D (1999) Shared information goods. J Law Econ 42(1):117–156

Boyer M (2018, forthcoming) The competitive market value of copyright in music: a digital gordian knot. Can Public Policy

Copyright Board of Canada (2002) Pay audio decision. http://www.cb-cda.gc.ca/decisions/2002/20020315-m-b.pdf

Lindahl E (1958) Just taxation – a positive solution. In: Musgrave R, Peacock A (eds) Classics in the theory of public finance. Macmillan, London, pp 98–123 (the original appeared as Chap. 4 Part I of Die Gerechtigkeit der Besteuerung, Lund 1919; translated by Elizabeth Henderson)

Osborne MJ, Rubinstein A (1994) A course in game theory. MIT Press, Cambridge

Supreme Court of Canada (2015) Canadian Broadcasting Corp. v. SODRAC 2003 Inc., 2015 SCC 57, File No. 35918

Varian HR (1999) Markets for information goods. Institute for monetary and economic studies, vol 99. Bank of Japan, Tokyo