Chapter 12
Norms in Game Theory

Davide Grossi, Luca Tummolini, and Paolo Turrini

12.1 Introduction

In this brief chapter we will overview several points of contact between games and norms. Since the following short exposition cannot be comprehensive, it aims to suggest a set of key ideas and perspectives connecting norms and games.

Generally speaking, the contributions in the literature at the interface between games and norms can be divided into two main branches: the first, mostly originating from economics and game theory (Coase 1960; Hurwicz 1996, 2008), exploits normative concepts, such as institutions or laws, as *mechanisms* that enforce desirable properties of strategic interactions; the second, that has its roots in social sciences and evolutionary game theory (Coleman 1990; Ulmann-Margalit 1977) views norms as *equilibria* that result from the interaction of rational individuals.

The chapter will reflect this division and be articulated in two parts. The first one—*norms as mechanisms*—will deal with those approaches within game theory (as well as related disciplines such as multi-agent systems (Shoham and Leyton-Brown 2008)) which study norms and institutions as components of games (e.g., mechanism design or implementation theory (Osborne and Rubinstein 1994,

D. Grossi (✉)
Department of Computer Science, University of Liverpool, Liverpool, UK
e-mail: d.grossi@liverpool.ac.uk

L. Tummolini
Institute of Cognitive Science and Technologies, CNR, Rome, Italy
e-mail: luca.tummolini@istc.cnr.it

P. Turrini
University of Luxembourg, Luxembourg City, Luxembourg
e-mail: paolo.turrini@uni.lu

S. Ossowski (ed.), *Agreement Technologies*, Law, Governance and Technology Series 8, DOI 10.1007/978-94-007-5583-3_12,

Chap. 10)). The second part— *norms as equilibria*—moves in the opposite direction reviewing approaches that use game-theoretic methods to explain and analyze norms, institutions[1] and their emergence.

12.2 Norms as Mechanisms

This section presents the view of norms as constraints that, imposed on players' behaviour, enforce desirable social outcomes in games. In this view, norms can be either seen as a way of engineering interactions from scratch, i.e., norms that dictate the 'legal' moves of a game, or as a way of transforming existing interactions, i.e., norms that modify the players' strategic possibilities in a game.

12.2.1 Norms as Rules of the Game: Mechanism Design

The view of norms as the rules of the game[2] is widespread within the so-called new institutional economics.[3] An interpretation of it from the standpoint of game theory is developed in Hurwicz (1996), which interprets the phrase literally in terms of the theory of mechanism design.

In brief, institutions are seen as collective procedures geared towards the achievement of some desirable social outcomes. An example of them are auctions, viz. mechanisms to allocate resources among self-interested players. In many auctions goods are not assigned to the bidder valuing them most as bidders might find it convenient to misrepresent their preferences. In such situations mechanism design can be used to enforce the desirable property of truth telling. For instance, when the bidders submit independently and anonymously and the winner pays an amount equivalent to the bid of the runner-up, truth telling is a dominant strategy.[4] In other words, in a second-price sealed bid auction, independently of the way they value the auctioned good, players cannot profitably deviate from telling the truth.

Viewing norms as mechanisms means considering them in the guise of auctions. Just like in auctions, they are supposed to make no assumptions on the preferences of the participating agents. They merely define the possible actions that participants can take, and their consequences. Slightly more technically, they are *game forms* (or mechanisms), viz. games without preferences.

---

[1] We will often use the terms "norm" and "institution" as synonyms.

[2] The phrase comes, as far as we know, from North (1990).

[3] New institutional economics has brought institutions and norms to the agenda of modern economics, viewing them as the social and legal frameworks of economic behavior. See Coase (1960) for a representative paper.

[4] This is the so-called Vickrey auction. See (Shoham and Leyton-Brown 2008, Chap. 11) for a neat exposition.

Two aspects of this view are particularly noteworthy. First, it clearly explains the rationale for norms and institutions: they are there in order to guarantee that socially desirable outcomes get realized (in jargon, *implemented*) as equilibria of the possible games that they support. Second, it presupposes some sort of infallible enforcement: implementation can be obtained only by assuming that players play within the space defined by the rules, which represents a strong idealization of the real workings of institutions.[5]

12.2.2 Norms as Game-Transformations

Norms can be conceptualized not only as the very framework of social interaction, like in the game-form conception above, but also as ways of *transforming* existing games in order to bring about outcomes that are more desirable from a welfaristic point of view. Game transformations include, for instance, appropriate restrictions of players' strategies or redistributions of such strategies among the agents.[6]

The game-transformation approach has been pioneered by Shoham and Tennenholtz (1995) in order to engineer laws which guarantee the successful coexistence of multiple programs. It has been further explored in the multi agent systems community to study temporal structures obeying systemic requirements, as in van der Hoek et al. (2007).

Sharing the same view of norms as game-transformations, the work of Grossi and Turrini (2010) investigates the role of interdependence in designing such norms. Instead of considering any arbitrary constraint on players' behavior, games are transformed respecting an underlying dependence structure among the players, i.e., taking into account what players would do if they could have a say on other players' actions. Inspired by previous work in social science (Castelfranchi et al. 1992), they also show formally how transforming games to implement desirable behavior is equivalent to enforcing a contract among the individuals involved, considering how players can mutually profit from one another.

12.3 Norms as Equilibria

Alternative to the view of institutions as 'rules of the game' is their conceptualization as equilibria, i.e., as stable behaviors, within games.[7] The difference might look subtle, but it is of a fundamental kind. Viewing institutions as game forms means viewing them as the 'hard constraints' defining the boundaries of possible

[5]This problematic assumption has been put under discussion extensively in Hurwicz (2008).

[6]See Parikh (2002) for an inspiring manifesto.

[7]This fundamental distinction has been emphasized, for instance, in Hurwicz (1996).

	C	D
C	2,2	0,3
D	3,0	1,1

	L	R
L	1,1	0,0
R	0,0	1, 1

Fig. 12.1 Prisoner's dilemma (with C = cooperate and D = defect) and Coordination game (with L = left and R = right)

interactions, while viewing them as equilibria means viewing them as some kind of 'softer' constraints from which it is possible, although 'irrational', to deviate.[8] Also, while the mechanism design view considers norms as an actual *component*—the game form—of the definition of a game, the equilibrium-based view considers norms as the result or *solution* of a game. So, in the former view norms define games, in the latter they are defined by games.

12.3.1 Social Norms

Starting from the classical problem of the spontaneous emergence of social order, the game-theoretic analysis of norms has focused in particular on informal norms enforced by a community of agents, i.e. *social* norms. From this perspective, the view of norms as Nash equilibria has been first suggested by Schelling (1966), Lewis (1969) and Ullmann-Margalit (1977). A Nash equilibrium is a combination of strategies, one for each individual, such that each player's strategy is a best reply to the strategies of the other players. Since each player's beliefs about the opponent's strategy are correct when part of an equilibrium, this view of norms highlights the facts that a norm is supported by self-fulling expectations.

However, not every Nash equilibrium seems like a plausible candidate for a norm. In the Prisoner Dilemma (see Fig. 12.1) mutual defection is a Nash equilibrium of the game without being plausibly considered a norm-based behavior. In fact, the view of norms as Nash equilibria has been refined by several scholars. Bicchieri (2006), for instance, has suggested that, in the case of norms conformity is always *conditional* upon expectations of what other players will do. Moreover, in this model, norms are different from mere conventions, in that norms are peculiar of mixed-motives games (e.g. the Prisoner Dilemma) and operate by transforming the original games into coordination ones.

Another influential view of norms characterizes them as devices that solve equilibrium selection problems. A comprehensive and concise articulation of this view can be found in Binmore (2007) which emphasizes two key features of norms.

[8] It might be worth stressing that the two views are not incompatible as institutions as equilibria can be thought of arising within games defined on institutions as game forms.

First, as equilibria, they determine self-enforcing patterns of collective behavior,[9] e.g., making cooperation an equilibrium of the (indefinitely iterated) prisoner's dilemma. Second, since repeated interaction can create a large number of efficient and inefficient equilibria, a norm is viewed as a device to select among them—a paradigmatic example of a game with multiple equilibria is the game on the right in Fig. 12.1, known as the coordination game.

Finally, it has been recently suggested that a norm is best captured as a correlating device that implements a correlated equilibrium of an original game in which all agents play strictly pure strategies (Gintis 2010). A correlated equilibrium is a generalization of the Nash equilibrium concept in which the assumption that the players' strategies are probabilistically independent is dropped. When playing their part on a correlated equilibrium the players condition their choice on the same randomizing device (Aumann 1987). Since the conditions under which a correlated equilibrium is played are less demanding than those characterizing Nash equilibria, the view of norms as a correlating device seems more plausible. Moreover, the correlating device is seen as a device that suggests separately to each player what she is supposed to do and thus seems to better characterize the prescriptive nature of norms (Conte and Castelfranchi 1995). On the other hand, the origins of such correlating devices is left unclear and are viewed as an emergent property of a complex social system.

Although an equilibrium-based analysis of norms might provide a rationale for compliance, it does not explain how such norms can possibly arise in strictly competitive situations—like the Prisoner's Dilemma. The next section discusses some approaches to this issue.

12.3.2 The Evolution of Norms

Axelrod (1986) studies norms starting from games in extensive forms with the following structure: the first player, i, chooses whether to comply or violate a (further unspecified) norm; if she violates it, a node is reached where nature chooses with what probability i's violation is observed by some other agent j; in case i's violation is observed, a choice node is reached where j has to decide whether to punish i or not; finally, the payoffs are the obvious ones for i, and j is assumed to incur costs when punishing i. In other words, the game provides a simple abstraction of norm compliance and defection, together with a basic enforcement mechanism. What Axelrod sets then out to do is to observe, by means of computer simulations,

[9] Self-enforcement is the type of phenomenon captured by the so-called *folk theorem*. The theorem roughly says that, given a game, any outcome which guarantees to each player a payoff at least as good as the one guaranteed by her minimax strategy is a Nash equilibrium in the indefinite iteration of the initial game (cf. Osborne and Rubinstein 1994, Chap. 8).

under what conditions[10] and how fast (i.e., after how many iterations of the game) compliance spreads among a population of players that randomly get to play role i and role j. In brief, the findings seem to show that, in order for compliance to arise, a meta-enforcement mechanism needs to be introduced, according to which j gets punished by other members of the population when not-punishing i.

A more analytical take on the evolution of norms can be found in Skyrms (1996), which uses techniques coming from the field of evolutionary game theory. The key idea behind this approach is to read games not as the interaction of players, but rather as the interaction of populations of strategies which are paired with each other; and payoffs not as utilities, but rather as the measures of fitness of the strategies that yield them. The higher the (average) fitness of a given strategy the larger will its population grow. The coordination game offers a very simple example: if the population of L (= drive left) strategies is more than half the whole population, it means that strategy L will have a higher average fitness than R, as R, under random pairing, will be more likely than L to end up in an uncoordinated outcome. Under this sort of evolutionary drive, the system will then stably reach the L,L equilibrium. As (Skyrms 1996) shows, this sort of analysis can be carried out to explain how equilibria are reached in all kinds of different games, and how even strictly dominated strategies (like cooperation in the Prisoner's dilemma) can be fixed into a stable evolutionary state.

References

Aumann, R. 1987. Correlated equilibrium as an expression of bayesian rationality. *Econometrica* 55: 1–18.

Axelrod, R. 1986. An evolutionary approach to norms. *The American Political Science Review* 80(4): 1095–1111.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.

Binmore, K. 2007. The origins of fair play. *Proceedings of the British Academy* 151: 151–193.

Castelfranchi, C., M. Miceli, and A. Cesta. 1992. Dependence relations among autonomous agents. In *Decentralized A.I.3*, pp. 215–227, ed. E. Werner and Y. Demazeau. Amsterdam: Elsevier.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* 3: 1–44.

Coleman, J. 1990. *Foundations of social theory*. Cambridge, MA: Belknap Harvard.

Conte, R., and C. Castelfranchi. 1995. *Cognitive and social action*. London: UCL Press.

Gintis, H. 2010. *The bounds of reason*. Princenton: Princenton University Press.

Grossi, D., and P. Turrini. 2010. Dependence theory via game theory. In *Proceedings of the 9th international conference on autonomous agents and multiagent systems (AAMAS 2010)*, ed. W. van der Hoek and G. Kaminka, 1147–1154. Richland: IFAAMAS.

Hurwicz, L. 1996. Institutions as families of game forms. *Japanese Economic Review* 47(2): 113–132.

Hurwicz, L. 2008. But who will guard the guardians? *American Economic Review* 98(3): 577–585.

[10]The conditions considered are essentially three: how risk-seeking is i, what the probability of a violation to be detected is, and how prone is j to react upon a detection.

Lewis, D. 1969. *Convention: A philosophical study*. Cambridge, MA: Cambridge University Press.
North, D. C. 1990. *Institutions, institutional change and economic performance*. Cambridge, MA: Cambridge University Press.
Osborne, M. J., and A. Rubinstein. 1994. *A course in game theory*. Cambridge, MA: MIT.
Parikh, R. 2002. Social software. *Synthese* 132(3): 187–211.
Schelling, T. 1966. *The strategy of conflict*. London: Oxford University Press.
Shoham, Y., and M. Tennenholtz. 1995. Social laws for artificial agent societies: Off-line design. *Artificial Intelligence* 73(12): 231–252.
Shoham, Y., and K. Leyton-Brown. 2008. *Multiagent systems: algorithmic*. Game-Theoretic and Logical Foundations. Cambridge, MA: Cambridge University Press.
Skyrms, B. 1996. *Evolution of the social contract*. Cambridge/New York: Cambridge University Press.
Ulmann-Margalit, E. 1977. *The emergence of norms*. Oxford: Clarendon Press.
van der Hoek, W., M. Roberts, and M. Wooldridge. 2007. Social laws in alternating time: Effectiveness, feasibility, and synthesis. *Synthese* 156:1: 1–19.