Gerhard Minnameier

The Economics of Morality and The Fabric of Social Sciences

Abstract: The paper presents an economic theory of morality and, based on it, a classification of social sciences and economic subdisciplines. Following the introduction, sections 2 to 4 seek to internalise morality into economics, both from a decision-theoretic and a game-theoretic point of view. In sections 5 and 6 moral principles are reconstructed as institutions, and it is shown why and how moral principles turn mixed-motive games into coordination games (along a hierarchy of moral stages and the respective games). This reveals the immense importance of morality not only in real life but also with respect to theoretical and empirical work in economics. Further, it may be asked whether there remains a proper realm for morality (or ethics) above and beyond the economic frame of reference. This issue is addressed in section 7, where the structural differences and the systematic relations between economic and neighbouring disciplines are discussed. Section 8 concludes and discusses important ramifications.

Keywords: Morality, Social Norms, Ethics and Economics, Game Theory

Zusammenfassung: Der Artikel stellt eine ökonomische Theorie der Moral und darauf aufbauend eine Klassifizierung der Sozialwissenschaften und ökonomischer Teildisziplinen vor. Im ersten Teil wird versucht, Fragen der Moral mit Mitteln der Ökonomik zu rekonstruieren, sowohl in entscheidungstheoretischer als auch in spieltheoretischer Perspektive. Der zweite Teil erörtert moralische Prinzipien als Lösungskonzepte für Mixed-Motive-Spiele. Dabei nehmen sie die Rolle von Institutionen ein und werden so ökonomisch internalisiert. Dies wirft wiederum die Frage nach einer „Demarkationslinie“ zwischen Ethik und Ökonomik auf, die im dritten Teil aufgegriffen wird, in dem die strukturellen Unterschiede und die systematischen Beziehungen zwischen den Wirtschaftswissenschaften und ihren Nachbardisziplinen (insbesondere der Ethik) erörtert werden. Der letzte Abschnitt schließt mit einem Ausblick auf wichtige Konsequenzen.

Schlagwörter: Moral, Soziale Normen, Ethik und Ökonomik, Spieltheorie

JEL classification: A12, D01, D90

Prof. Dr. Gerhard Minnameier, Goethe-Universität Frankfurt, FB 2: Wirtschaftswissenschaften, Theodor-W.-Adorno-Platz 4, D-60629 Frankfurt am Main, E-Mail: minnameier@econ.uni-frankfurt.de

https://doi.org/10.1515/ordo-2025-2037

Content

1. Introduction: The quest for morality from the economic point of view

Morality seems to be multifaceted and elusive. We think of it in terms of internalised values and principles, i.e., the "moral point of view", but also in terms of rules that govern social life and enable human cooperation. Some think morality consists mainly in reasoned judgements (Kohlberg 1981; Gert 2005; Scanlon 2014); others hold that moral reasoning is mostly an ex post rationalisation of what has already been decided emotionally or intuitively in "system 1" mode (Haidt 2007; Greene 2013).[1]

While the message before and after the turn of the century was that real humans had "social preferences" that had to be built into economic models (Fehr & Schmidt, 2006), views have somewhat changed, after it had become clear that those preferences are mostly crowded out by anonymity or moral wiggle room (Hoffman, McCabe & Smith 1996; Dana, Weber & Kuang 2007; Andreoni & Bernheim, 2009) as well as by changes in choice sets (List 2007; Bardsley 2008). Dana, Weber and Kuang got to the heart of the ensuing change of mind when they stated that "(r)ather than having a preference for a fair outcome, people may conform to situational pressures to give in certain contexts, but may also try to exploit situational justifications for behaving selfishly" (2007, 69). More such results and interpretations have been published under the label of "moral hypocrisy" (Batson et al. 1999; Lönnqvist, Irlenbusch & Walkowitz 2014; Rustichini &

1 What's more, when people decide spontaneously, they usually take the moral course of action; however, when they reason deeply, they tend to take self-interested decisions (Cappelletti, Goth & Ploner 2011; Rand, Greene & Nowak 2012; Greene, 2013, 62–63).

Villeval 2014). However, if this were true, one might wonder why moral rules exist at all and why, as it seems, many "good people" fail miserably at the simplest moral tasks.

Contrary to the pessimistic interpretation just mentioned, there has also been a long history of research in economics and the related field of ethics – mainly within the German-speaking context – which points to situational constraints as the root causes of moral failure (e.g., Brennan & Buchanan 1985; Buchanan 1977; Binmore 1994; Homann & Bloome-Drees 1992; Homann & Pies 1994; Homann & Suchanek 2005; Lütge 2005; 2012; Lütge & Mukerji 2016; Lütge & Uhl 2018; Minnameier 2012; 2013a, b; 2020; Pies 2009). To be sure, however, these authors claim that there is no real moral failure, since it is the institutional framework that is to blame, not individual agents.

Questions about moral functioning are of interest not only to those working in the field of social preferences but also in the wider economic perspective. Ever since James Buchanan asked "What should economists do?" (1964) has it been clear that (modern) economics is basically a science concerned with human cooperation, broadly considered. From this viewpoint, morality in all its various forms seems to be an important and powerful economic tool. The present paper tries to reveal just how it works, both in principle and in systematically different contexts.

The first part seeks to internalise morality into economics, first from a decision-theoretic point of view (section 2), then from a game-theoretic point of view (section 3). The second part discusses moral principles as institutions and explains how moral principles turn mixed-motive games into coordination games (section 4) and how a hierarchical order of such games gives rise to a hierarchy of moral principles as solution concepts (section 5). This internalisation of morality into economics motivates the question about what remains for other disciplines, ethics in particular, and how to map the transdisciplinary field in terms of different kinds of approaches to one and the same issue, which is addressed in the third part (section 6), where the structural differences and systematic relations between economic subdisciplines and neighbouring disciplines are discussed. The final section concludes and discusses important ramifications.

2. Morality in terms of social preferences

If we want to understand morality in terms of social preferences, we typically take on a decision-theoretic perspective, from which other individuals are conceived of as parts of the agent's environment. Their preferences and beliefs belong to the overall set of restrictions, under which the agent forms an intention. These restrictions can be positive (i.e., affordances) or negative (i.e., constraints), and intentions can be formed deliberately or intuitively. A rational moral choice would then be utility-maximising regarding the agent's (social) preferences and the restrictions under which she chooses.

However, we have to distinguish between two completely different notions of "preference". The first is the concept of "revealed preference" which pertains to consequences of choices (Samuelson, 1948). The second is the concept of underlying *funda-*

mental preferences that motivate *concrete preferences* (over consumption bundles). For instance, a preference for a small car – if a larger one were affordable – might be motivated either by thriftiness or by a concern for the environment (Dietrich & List 2013a, b). In the moral domain, principles like the "Golden Rule" might be motivating reasons for the proposer in the dictator game to prefer a fair outcome or for the trustee in a trust game to transfer a fair share back to the investor.

It is important to note that these fundamental preferences are out of reach for the strictly behaviouristic concept of revealed preferences, which typically relates to observable outcomes of choices. In this sense, rational choice theory based on "revealed preferences" has been criticised for depriving utility of all content (Hollis & Sugden 1993; Bruni & Sugden 2007; Hausman 2012). Conversely, the idea of fundamental preferences (re-)endows utility with content. It brings the central drivers of choices to the fore, among them also the ones that we generally conceive of as "social preferences".

While this approach has much in common with the psychological approach originally set forth by Becker (1976), it also differs importantly from it. Becker's notion of "basic preferences" refers to motivators common to all humans, much like the basic needs assumed by Deci and Ryan in their self-determination theory (2017). In contrast to this, fundamental preferences as understood here may vary inter-individually (see also Dietrich & List 2017). Vanberg has discussed Becker's approach extensively and argued that the economists' reluctance to introduce such fundamental preferences is unjustified (1988/1944, 42–50). Moral preferences also differ from social preferences. The latter are modelled as preferences over outcomes or, in a more elaborated version, about the intentions interaction partners (see e.g. Meier 2007 for an overview). However, Vanberg (2008) argues that they ought to be understood as preferences over actions, which are characteristically different in that morality is about doing the right thing rather than bringing about a certain outcome (although both aspects are clearly related to each other) (see also Hodgson, 2019, 118–123). Accordingly, we can analyse moral agency in a choice-theoretic frame of reference, where a moral person has fundamental preferences in terms of her deep moral convictions but faces certain restrictions in a specific situation. And a morally rational choice can be determined for agents with clearly defined moral preferences (e.g., Minnameier 2016).

It can also be shown that different sorts of moral preferences – in terms of moral stages – are activated in response to specific situational constraints. For instance, a person who tries to follow the *Golden Rule* ("Do unto others what you would have others do unto you") might be motivated to cooperate in a one-shot prisoner's dilemma (PD), since she would not wish her partner to defect. However, given the risk of defection, this strategy is self-defeating, which explains why not only self-interested but also morally motivated people learn to defect (see Ledyard, 1995). Hence, the one-shot PD is a situation that requires another kind of morality, which is "Every man for himself" (Minnameier 2013b).

In moral psychology it has long been clear that people follow different moral guidelines in different situations (e.g., Beck, Dransfeld, Minnameier & Wuttke 2002;

Krebs & Denton, 2005; Minnameier, Beck, Heinrichs & Parche-Kawik 1999; Rai & Fiske, 2011), because different situations call for different kinds of morality. Thus, the crucial question is not: "Which model of other-regarding preferences does best in the light of the data ...?" (Fehr & Schmidt, 2006, p. 668) but how to distinguish *different* types of fundamental moral preferences and the conditions of their proper use in terms of a rational choice accounting for situational constraints. Hence, reconceiving social preferences in the sense just described allows us to reconstruct moral agency in many different situations and also to evaluate what morality commands or does not command in order to prevent the exploitation of this very morality.

3. Morality in terms of "rules of the game"

However, reconceiving social preferences in terms of fundamental moral preferences may still not tell the entire truth. As mentioned in the introduction, various experiments have revealed that most participants (in general 70 per cent or more) fail to follow what could be called the moral course of action (Andreoni & Bernheim 2009; Batson et al. 1999; Dana, Weber & Kuang 2007; Hoffman, McCabe & Smith 1996). They all concern variations of the dictator game, in which the characteristic features are (1) that players are anonymous or cannot be made accountable and (2) that the players making an active choice (i.e., the dictators) face no restrictions that would rationally prevent them from following the moral course of action. The latter is the key point: A dictator can always determine a fair outcome without any self-harming consequences apart from the foregone profit. Therefore, it seems hard to reconcile this with the idea that most people actually do have (strong enough) moral preferences.

Consider, e.g., the phenomenon of moral hypocrisy. In Batson et al.'s (1999) setting, the active agents[2] typically choose to flip a coin to determine how to assign different tasks (one favourable, one unfavourable) to another person and to herself. They are not observed when flipping the coin, and a moral agent can just follow the result of the coin flip, yet 90 per cent of those flipping the coin report that the coin decided in their own favour. It seems that people (mis)use moral wiggle room for downright selfish reasons. However, it does not imply that those people act immorally, as I try to show in what follows.

From a theoretical point of view, we have to recall the two different notions of morality already discussed above. So far, we have considered moral principles in terms of other-regarding preferences. Binmore has criticized this approach for "blurring the distinction between a social norm and a social preference" (2010a, 141) and believes this

2 Other than in economic experiments there are no real passive players in this social-psychological experiment, but the active participants belief that there is another real person who will have to actually carry out the assigned job.

"will turn out to be a bad mistake" (ibid.), because it treats a game-theoretic problem as a decision-theoretic one.

However, if moral principles function as social norms, they are to be understood as institutions. This is the view that Binmore (2010b) himself advocates. However, he also raises the question of whether social norms (or moral principles) can really be rules of the game in the game-theoretic sense, because for this they would have to be unbreakable. If rules were implemented by some powerful and incorruptible external ruler, they might be unbreakable for the players as a matter of fact. However, if moral principles are simply rules in terms of socially shared views about what one ought to do in certain situations, agents could always violate them. And if they can be violated, they cannot possibly be rules of the game in the game-theoretic sense. Binmore goes on to argue like this:

> The game theory solution to the problem that arises when the rules of an institution are not enforced by some incorruptible external agency is to move to a larger game—the game of life—in which the institution is regarded as being embedded. This larger game must have rules that are genuinely unbreakable, like the laws of physics, because if the players had strategies whose implementation resulted in the rules being broken, then they would not be the genuine rules of the game. The rules of the institution to be studied then have a lesser status. The players can break them if they want to, but if the institution is stable, the rules will not be broken, because obeying them is part of the behavior required by an equilibrium of the game. That is to say, an institution is not treated as a game itself, but as part of the description of an equilibrium within a larger game of life" (Binmore 2010b, 246).

Hence, in the larger game of life, the players may discover an equilibrium in which they can coordinate. If they succeed in doing so, they will implement the institution as a rule for the (smaller) base game, where it will not be broken if the players are in their right mind. (see also Greif & Kingston 2011) In equilibrium, they act as (collaborative) "choreographers" who redesign the base game in which they act as ordinary agents (cf. also Aoki 2011; Guala 2016; Guala & Hindriks 2015; Vanberg 1988/1994). This also allows for developmental progress in terms of an emerging order of games, which extends and complements Hayek's notion of spontaneous order (Vanberg 1986/1994; Vanberg & Buchanan 1988/1994).

Now, if moral principles function in this way, i.e., as institutions, they have to come with sanctions to enforce them and to make cooperation the dominant strategy. Only few have come to understand moral principles themselves as institutions or as social norms in the economic sense. The first was probably Edna Ullmann-Margalit (1977), but also authors like James Buchanan and Viktor Vanberg have been among the pioneers in this respect. Both have identified social dilemmas as the underlying problem and morality as the solution, also with respect to the sanctioning potential of reputational information spread in response to moral behaviour or misbehaviour (see e.g., Vanberg 1988/1994, 51–54; Vanberg & Buchanan 1988/1994). More recently, Cristina Bicchieri (2006; 2017) has discussed norms as solutions to mixed-motive games in Schelling's (1960) sense. If she is right, morality has a proper place in economics, because moral principles would belong to the toolbox we use to make cooperation possible (see also

DeScioli & Kurzban 2013; Guala 2016; Sugden 2018, chap. 11). If Buchanan is right in claiming that economic problems are typically problems of human cooperation, the relevance of morality for economics can hardly be overestimated (see also Binmore, 2011, p. 171).

Let us return to the question of how moral rules can be institutions, and let us consider the simple example of the Hawk-Dove Game. In Hawk-Dove, players compete over a resource of value v and can either play "hawk" (H), which means to fight until one either wins the resource or is injured and has to retreat (with cost of injury c), or "dove" (D), which is to display hostility but retreat before any kind of fighting would actually ensue. Let $v = 20$, $c = 40$, and let us further assume that both combatants have equal strength, so that there is a 50 percent chance of winning for each of them. Fig. 1 shows the resulting strategy profiles and the payoffs.

	D	H
D	*v/2, v/2*	*0, v*
H	*v, 0*	*(v-c)/2, (v-c)/2*

a

	D	H
D	*10, 10*	*0, 20*
H	*20, 0*	*-10, -10*

b

Figure 1: The hawk-and-dove game with (a) the payoffs in general form and (b) the payoffs if $v = 20$, $c = 40$

Unlike the PD, HD does not have a stable pure strategy Nash equilibrium. There are two pure strategy Nash equilibria (H, D and D, H), which, however, are asymmetric. On top of these, there is a symmetric equilibrium in mixed strategies, in which each player chooses H and D with probability $p = .5$. In this case, each player earns an expected payoff of 5. However, similar to the PD, this symmetric mixed strategy Nash equilibrium is Pareto-inefficient, because both players end up with *5*, when 10 would have been possible (D, D).

Fortunately, there is a way out – not within the game, but by augmenting and thereby changing it. This is done by introducing a new rule, the so-called "property rule" (P), which turns HD into a new game ("Hawk-Dove-Property", or HDP) (Sugden, 2005, pp. 73–74; 2010, pp. 51–52; Gintis, 2014, pp. 145–146). The rule introduces a new strategy P, namely, "When first at the resource, choose H, otherwise choose D." This is a realistic and common rule. Children learn it very early, e.g., when they compete over toys at a nursery school. Adults use it when competing over parking spaces or over seats on trains, or in desk-sharing offices. What's more, this is not just any rule, but clearly a moral rule.

Let us now look at this new game, HDP, and let us further assume equal chances to be the first (or second) at the resource. Under these conditions, we obtain the following payoff matrix (see Fig. 2). For reasons of simplicity, only the row-player's payoffs are shown.

	P	D	H
P	*v/2*	*3v/4*	*(v-c)/4*
D	*v/4*	*v/2*	*0*
H	*3v/4 – c/4*	*V*	*(v-c)/2*

A

	P	D	H
P	*10*	*15*	*-5*
D	*5*	*10*	*0*
H	*5*	*20*	*-10*

b

Figure 2: The Hawk-Dove-Property game with the payoffs for the row-player: (a) in general form and (b) if $v = 20$, $c = 40$

If the column-player chooses P, the row-player's best response is to choose P, too. Hence, the payoffs (10, 10) are now attainable. Moreover, whereas HD has the structure of a "mixed-motive game" according to Schelling (1960), the introduction of P has transformed the game into a coordination game. Extending this basic insight, the central claim of this paper is that moral principles generally function as solution concepts for mixed-motive games that transform them into coordination games, where the Pareto-superior strategy profile is a Nash equilibrium.

Not only must the moral rules be common knowledge, but the players also have to be able to use the sanctioning mechanisms that are part of the game. To see this, let us assume the column-player chooses strategy D because she has been educated to be kind to others. In the case of HDP, this corrupts morality, because it amounts to an invitation for the row-player to choose H. Hence, the important lesson from this is that if moral principles are institutions, the respective sanctions have to be employed appropriately.

4. Correlated equilibria and a hierarchical order of games

Correlated equilibria, originally a concept of cooperative game theory, were introduced by Robert Aumann (1974) in the context of non-cooperative game theory. From the point of view of cooperative game theory, they determine only what would be beneficial (the "efficient" outcome) for a coalition of players, not how they might achieve it. However, as Gintis stresses, "the correlated equilibrium is a much more natural equilibrium criterion than the Nash equilibrium, because of a famous theorem of Aumann (1987),

who showed that Bayesian rational agents in an epistemic game *G* with a common subjective prior play a correlated equilibrium of *G* (Gintis 2014, 142; see also Vanderschraaf, 1995; 2019, 60–62). This raises the question of how they manage to coordinate in this way. Of course, the straightforward answer is that they must do something so that the correlated equilibrium becomes also a Nash equilibrium, such as the move from HP to HDP.

Such a move is not part of classical game theory, and it cannot even be part of game theory in a narrow sense, because it pertains to transforming or inventing games. Changing the (rules of the) game is never a strategy within the respective game; it would only be one in the larger game of life. From the point of view of classical game theory, this would require omniscient players. However, from the point of view of evolutionary game theory, it does not, because players can find equilibrium paths intuitively (Binmore 2010b; 2011). However, we can also imagine the players themselves establishing new rules and subsequently following them (Fudenberg & Tirole, 1991, p. 53; Vanberg & Buchanan 1988/1994; see also sec. 3).

In HDP, players can thus establish the "property" rule, which allows them to achieve the correlated equilibrium, as it is now a Nash equilibrium. Of course, this does not constitute property in a formal sense, but property in the sense of legitimate possession, like the possession of an unreserved seat on a train that one has just occupied.

However, within this new game, new cooperation problems may arise. Once property is established, it might happen that player A lacks what player B has and vice versa. Sharing and turn-taking are the simple moral rules that solve such problems. Children learn to share their property with their friends and take turns on playground equipment. Office workers take turns using common resources such as specialised software or office workplaces. Thus, the sharing norm is the moral principle for circumstances in which not everyone has his or hers, i.e., where the property rule as such does not suffice to solve the problem. Hence, we end up with a more complex game, because the sharing norm implies the property rule but goes beyond it.

Finally, yet another problem arises whenever the players are relevantly different from each other. For instance, if one of the office workers has a very important job, she might skip the queue, or when on a bus, tram or train, one might give one's seat to an elderly or disabled person. Other relevant differences concern effort and expertise. Sales representatives and many other employees receive a basic remuneration plus a premium depending on sales or other measures of success, so that the overall payment to the employees is divided equitably rather than equally. In all such cases equal sharing would be unfair. Principles of care and equity account for special needs and achievements and help us to establish equitable distributions.

If we extend these basic ideas and try to reframe them in the context of Binmore's notion of a "larger game of life", institutions such as the ones that have been discussed so far allow agents to solve mixed-motive games so that they manage to coordinate in correlated equilibria in more and more complex social contexts. Moreover, these correlated equilibria are indeed Nash equilibria in the larger game of life. Thus, one

solution to a mixed-motive game may induce new kinds of such games at higher orders. And all these solution concepts, taken together, may constitute a hierarchy of moral principles that shape and govern a hierarchy of embedded games.[3]

In order to illustrate this idea, let us change the usual presentation of payoff matrices in such a way that the Pareto-inferior Nash equilibrium appears in the south-westerly cell and the correlated equilibrium in the north-east (see Fig. 3),[4] and let us use the PD as the paradigmatic example of mixed-motive games in general. If the prisoners could talk to each other, they could solve the PD by agreeing on a mutual promise together with suitable sanctions (any kind of credible threat to ward off defection). The sanctioning mechanism discounts the value of unilateral defection by three points. This move transforms the PD into a coordination game.

	D	C
C	*0+3, 4-3*	*3, 3*
D	*2, 2*	*4-3, 0+3*

Figure 3: PD transformed into a coordination game by implementing a mutual promise with sanctions

If we construct games at different emergent orders, as suggested above, we obtain an order of games at different (ordinal) welfare levels (see Fig. 4). With the move from one game to another, a new problem in terms of a new mixed-motive game emerges, based on the solution of the previous one.

3 The hierarchy of games discussed here and explained as well as exemplified in the sections 5 and 6 differ from the ordonomic three-stage model (see e.g. Pies 2000, 189; 2022, 29). Other than Pies, I have argued for reinterpreting the rationale of institutional change in the sense of the original two-stage logic of playing by the rules a game and changing the rules of the game (Minnameier 2025). Another hierarchy discussed by Pies in the appendix of his (2022, 219–221) does not concern a hierarchy of games in the sense of moral stages because what he exemplifies and discusses is merely an escalation of sanctioning mechanisms that do not change the moral framing.

4 This ingenious move to flip the payoff matrix so that it can be mapped onto a diagram to compare the agents' utilities was – as far as I know – introduced by Ingo Pies within his ordonomic framework (2000; 2009; 2016; 2022).

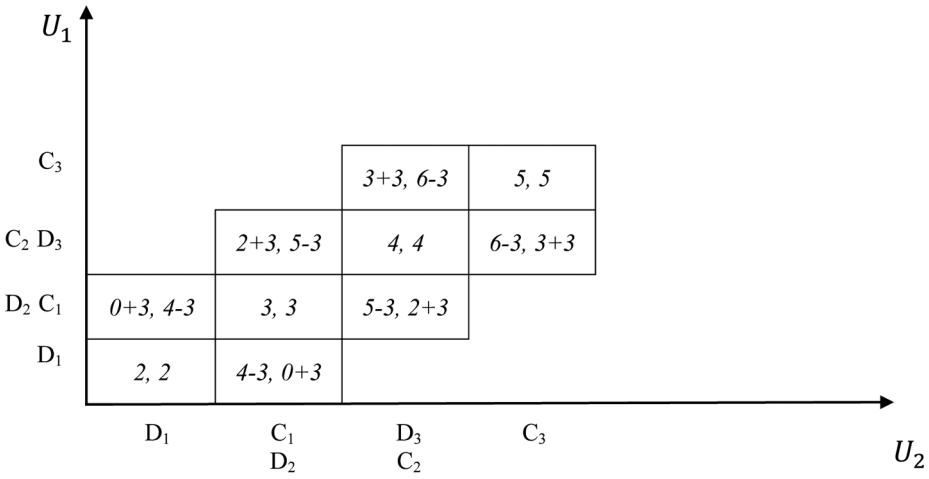


Figure 4: A succession of three (successively transformed) PD-like games at different levels (as indicated by the index numbers of the strategies). The ordinal numbers in the boxes signify that at each higher level higher payoffs are possible for the players. So, once the first game is transformed, players will reach the Pareto-superior combination (C1, C1) because it is a Nash equilibrium, but they also become aware of the next-higher game and realise that (4,4) were possible. However, they can only attain this payoff combination by transforming this second game, too, and so on.

5. Moral principles as institutions and the theory of moral stages

For moral principles to function as institutions, they have to come with sanctions. What kinds of sanctions might these be? Experimental economics usually operates with monetary incentives (Croson 2005), which do the job in the laboratory but may be counterproductive in real life (Bowles 2016; Gneezy & Rustichini 2000; Mulder et al., 2005; 2006; Mulder, 2009)[5] and obscure the functioning of subtle moral rewards and punishments in many forms of human cooperation (see also Kamenica 2012).[6] Generally, we seem to be so habituated to playing moral games in everyday life that we play them

5 However, based on the moral theory presented here it seems that material incentives do not crowd out morality altogether, but rather act as game changers in the sense of replacing one kind of morality for another one. For instance, in the classic example, where parents pick up their children late from a nursery school, an induced monetary punishment is (mis)understood as a price, which has led to more late pickups rather than less (Gneezy & Rustichini 2000). That is, an appeal to decent behaviour (Stage 3B) or the golden rule (Stage 2C) turns into kind of deal (Stage 2B) (see stage descriptions below).

6 Non-monetary payoffs have so far been analysed in terms of social preferences only, but not in terms of incentives. Nonetheless, monetary incentives may convey non-monetary signals. In a public goods game, e.g., they not only reduce the punished player's payoff, but may also signal indignation or the like.

as if they were downright coordination games (Bicchieri 2006, p. 41) and fail to notice the underlying dilemmas. But when morality is crowded out, those dilemmas re-emerge (see also Vanberg 1988/1994, 56–59).

Although some have understood and highlighted the role of non-monetary and specifically moral sanctions (Bicchieri 2006; Buchanan, 1994[7]; 1988/1994; Vanberg/Buchanan, 1988/1994), none of these authors has revealed how these sanctions work precisely; see, e.g., Lütge's (2012, 63–68) critique of Buchanan (1994) in this respect. Accordingly, it has so far been an open and largely neglected question as to what different kinds of moral sanctions would have to be distinguished and what different kinds of moralities, in terms of moral stages or something similar, would have to be differentiated. In what follows I will therefore try to reveal how moral sanctions actually function with respect to specific moral principles. In particular, we can differentiate three characteristic "moral currencies" that apply to three basic moral stages (see Tab. 1):

(1) *Sympathy vs. antipathy,* which operates where the involved agents have an intrinsic interest in each other, enjoy or aim at social bonding and therefore have a direct interest in the other's well-being.
(2) *Respect vs. contempt,* which operates in cases of conflicts of interest, where agents do not have an intrinsic interest in each other or where acting in the interest of the other directly impedes one's own.
(3) *Repute vs. disrepute,* which operates within a group or a community of different groups and where cultural rules of decency and role responsibilities are relevant.[8]

The stages as such are the result of a post-Kohlbergian approach that tries to build on the basic and still valid foundations of the well-known Kohlberg theory while it overcomes its inherent problems and shortcomings (see Colby & Kohlberg 1987; Kohlberg 1984, for the original theory, and Minnameier 2000; 2001; 2014, for the critique and the foundations for a revised theory of moral stages). The core of Kohlbergian theory is that there is a hierarchy of stages (of successively higher complexity) as illustrated above, where the higher stages integrate and extend the lower ones, and where a higher stage emerges as a result of (new and inherent) problems incurred at the preceding one.

Kohlberg differentiated six stages at two main levels and later added sub-forms (A and B) to each stage as well as transitional stages (especially 1/2, 2/3, 4/5). The post-Kohlbergian taxonomy encompasses nine stages with three substages each (A, B, C). However, only the first three stages are relevant here and will be described briefly below. By and large the post-Kohlbergian stages correspond to the original Kohlberg stages, although Stage 1 is conceived differently and the substages of Stage 2 and Stage 3

7 See also Carden, Caskey and Kessler (2022).

8 Reputation in the sense used here is narrower than the concept generally used in economics. Economists tend to call any belief one individual holds about another as a reputation of that individual. However, here it refers to socially shared beliefs that are formed and communicated a group or community by way of gossip or similar means.

do not always match perfectly with Kohlberg's stages: The Golden Rule, e.g., is associated with Kohlberg stage 3, whereas here it is classified as Stage 2C.

The sequence of substages A through C follows three basic justice operations[9] that recur at each stage (while each stage introduces a new relevant aspect that has to be integrated into moral judgement). The first, *equality*, means that each of the concerned persons' (or groups') individual perspectives are addressed and taken into account in the same way. The second, *reciprocity*, links these perspectives in a strictly reciprocal way in the sense of establishing an equal exchange balance between the respective agents. However, since such an equalised transfer balance may be unfair if the agents are characterized by relevant inequalities, the third justice operation, *reciprocal equality*, has to take account of these inequalities. To put it differently: The first operation allows us to address different individuals' respective viewpoints from an intra-individual perspective, the second allows us to relate them from an inter-individual perspective, while the third allows us to integrate them from an overarching trans-individual perspective. Figure 5 illustrates the justice operations and their role in moral development, in particular that with each stage a new aspect is introduced, under which perspectives are first differentiated, then reciprocally related, and finally integrated. For examples see section 6.

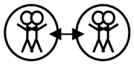
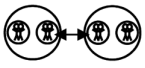
Stage type:	Intra	Inter	Trans
Stage 1: *Substages A – C*	A 	B 	C 
Stage 2: *Substages A – C*	A 	B 	C 
Level 3: *Substages A – C*	A 	B 	C 
Justice operations:	Equality	Reciprocity	Reciprocal equality

Figure 5: Illustration of the justice operations and their role in moral development

9 It should be noted that Kohlberg introduced the idea of justice operations (also relating to Piaget) in the context of his stage theory. However, the operations he mentions, i.e. equality, equity, role-taking, universalizability, sympathy and liberty) not only appear unsystematic in that it is unclear in what sense they really are operations. He also holds that they apply to each stage, which blurs their proper role for particular stages (see Colby & Kohlberg, 1987, 23–25; Kohlberg, Levine & Hewer, 1983).

Tab. 1 presents the three stages together with the relevant rewards and punishments in moral currency and the justice operations. Rewards for complying with morality increase utility (e.g., signals of affection or gratitude in the case of Stage 1), whereas punishments decrease utility (e.g., withdrawal of love in a partnership or family relationship at Stage 1). On top of this, there exists an extended form of punishment in case one's opposite number fails to react in response to punishment in moral currency. There is always the possibility of stopping to play this game altogether and moving to a lower game. In the case of the sharing norm (1B), this means to revert to a game in which nobody shares (anymore) but sticks to what they have (incorporating a moral game according to Stage 1A). This extended form of punishment might function as an incentive to get the other agent back on the path of virtue, so to speak. From the point of view of the lower-level game of property, a similar strategy could be employed: If the other player (continually) violates my property, I could first get angry and express my indignation, but the ultimate move would be to start conquering back items, which would be tantamount to reverting to a Hobbesian state of nature, in which, according to the "law of Nature (...) every man has right to everything" (1651/2001, p. 65 [chap. 15, § 2]), but which results in a "war of every one against every one" (1651/2001, p. 59 [chap. 14, § 4]).

Table 1: Moral principles and sanctioning mechanisms

Stage	Moral Principle	Reward	Punishm't	Justice Op.
3C	Legal norms (authority)	Repute	Disrepute	Recipr. Equ.
3B	Cultural norms (decency)	Repute	Disrepute	Reciprocity
3A	Group norms (role obligations)	Repute	Disrepute	Equality
2C	Golden rule (consideration)	Respect	Contempt	Recipr. Equ.
2B	Mutual promise (contract)	Respect	Contempt	Reciprocity
2A	Legitimate (competing) interests	Respect	Contempt	Equality
1C	Care (based on needs and merit)	Sympathy	Antipathy	Recipr. Equ.
1B	Sharing/turn-taking	Sympathy	Antipathy	Reciprocity
1A	Property (legitimate possession)	Sympathy	Antipathy	Equality

6. Description of the moral stages and substages

Moral Stage 1

The examples I have mentioned in sections 4 and 5 pertain to simple principles of morality that typically apply to close relationships based on feelings of sympathy.[10] In such situations, the players *do have an intrinsic interest in each other* and in each other's well-being. Therefore, the moral currency of this stage consists in signals of "sympathy" and "antipathy", as explained already in the previous section.

It has long been known that the three moral principles of the substages to Stage 1 follow a developmental order. In particular, Damon (1977) found that early preschool children behave selfishly in the sense of not sharing (strict property orientation; Stage 1A), that later preschool children share equally, and that at school age they share equitably. There is also recent empirical evidence for this developmental order from preschool children's sharing behaviour (see e.g., Paulus and Moore 2014; Killen, Elenbaas & Rizzo 2018; Smith & Warneken 2016).[11] However, for those moralities to be upheld, they have to be Nash equilibria in the larger game of life. Their functioning relies on (mutual) sympathy and a concern for the other's well-being (see also Paulus & Moore, 2017), and in that sense they create and promote friendship and affection.

Such a caring and sympathetic attitude may also explain why Casal, Fallucchi, and Quercia (2019) find a high proportion of altruism towards the (passive) third party in a three-player ultimatum game, where the third party is an NGO whose money can be taken away and redistributed between the other players. This is true for proposers, of whom one third refrain from taking money from the NGO (even where there is no risk involved), and for responders, who punish proposers for taking from the NGO (ibid., p. 77).[12]

10 The property norm, however, can be interpreted also in terms of higher forms of morality. The full notion of property in terms of permanent ownership and rules concerning its acquisition belongs to a higher form of social rules that govern life in (large) communities (see Stage 3 below).

11 Hamann, Bender and Tomasello (2014) also found meritocratic sharing in preschool children that preschool children, however only in a very limited form. First, it only occurred when children collaborated, not when they worked in parallel. Second, the meritocratic sharing consisted in a tendency to share equally, after rewards were given undeservingly unequally (a child who had to work more than the other received one marble, while the other received three). Third, the induced difference that was thought to bring about differential deservingness was only that one child had to work harder to complete a certain task, but which was completed together (so that the advantaged child could actually not be more productive than the other child). In my view, the result can be interpreted as a combination of claiming one's property (that has been allocated by the experimenter) and the equal sharing norm, but not a full account equitable sharing.

12 Furthermore the authors suggest that there are two moral types, because apart from the altruism which seems to motivate some players, other players are motivated by inequality aversion. While the former is a case of Stage 1C, the latter can be understood as implementing a fair deal in the sense of Stage 2B (see below).

Moral Stage 2

While Stage 1 is based on sympathy and genuinely common interests, there are also true conflicts of interest in which helping others may be fatal. Consider, for example, the situation of a couple of graduates applying for jobs over which they (have to) compete. They may feel a lot of sympathy for each other and might be friends, yet they clearly are competitors because acting in the interest of the other would compromise their own legitimate interest. In such a conflict of interest, the agents have to follow the rule that everybody has interests to pursue, which constitutes Stage 2A (see Tab. 1). Or in a proverb: *"Near is my shirt, but nearer is my skin."* As a principle, this does not sanction or even promote selfishness because it necessarily applies to all other agents as well and therefore implies respect for their pursuing *their* interests.[13]

This kind of reasoning also extends the moral scope in that it covers interactions with people to whom one might be indifferent or even opposed. Market interaction in particular leads to respectful and fruitful interaction between strangers and also people of different value systems. This applies especially to Stage 2B, where the agents cooperate in mutual exchange, while each of them seeks their own benefit.

Thus, Stage 2B relates to conflicts of interest that cannot be solved by just having everybody pursue their interest on their own, as at Stage 2A. The classical one-shot PD is another case in point. Of course, in a one-shot PD the social dilemma is irresolvable, and therefore the only rational option is to follow the principle of Stage 2A, which forces players to defect. However, as mentioned above, under more relaxed circumstances in which communication is possible, the morality of contract in terms of a mutual promise allows agents to solve the problem. To be sure, however, the sanction does not necessarily have to take the form of a material threat but can consist of a loss of respect.

In orthodox game theory this would count as cheap talk. However, for moral persons this talk is not cheap. Compare the situation of the strict PD with the one in which a promise was given. In the first case, no one can reproach the other for defecting (since it perfectly complies with the morality of Stage 2A). However, if a promise was first given and then broken, the other may rightfully hold the defector in contempt. And the defector will know this and have a bad conscience (given the respective sense of morality).

Finally, Stage 2C pertains to the Golden Rule, which is required in conflicts of interest in which the other party has nothing to offer in return for some service. In this situation, one is unable to strike a deal but rather has to make a deal with oneself in terms of what one would wish if one were in the other's shoes. The rationale behind this is that in some cases one is in a position to help; in other cases, one needs to be helped. Knowing the latter may lead to the conviction that being helpful and thus respectful

13 The well-known "Battle of the Sexes" would also be a case in point, if going to the theatre with his wife would not be better for the husband than going to the football match alone (and vice versa). In this modified game everyone would have to respect the other's interest, and everyone would have to go on their own (with whoever), or else they might either have to divorce, in the long run, or give in to the other.

towards others is in everybody's interest and should therefore be accepted as a general moral duty (at least in relation to those who think and feel the same). The trust game (see Johnson & Mislin, 2011), as well as a more sophisticated version of it (Hoffman, Yoeli & Nowak 2015)[14], both incorporate this very idea. Here, the investor must have reason to believe that the proposer plays according to the Golden Rule. Only if this condition is met can this moral game be played and can the players obtain a fair and efficient payoff.

In an experiment based on an augmented trust game, we could show that the possibility to exchange signals in moral currency increases both trust and trustworthiness substantially (Bonowski & Minnameier 2022). In post-play communication players could evaluate the other player's choice of action (in particular whether it was fair or unfair on a five-point Likert scale). Since it always happened after the player's move and players remained entirely anonymous, this kind of communication had no strategic value for monetary payoffs, yet it stabilised both trust and trustworthiness even across a succession of ten rounds (although pairs were matched anew after each round).

Moral Stage 3

Even when following the Golden Rule, you cannot always please everyone. For instance, a sales representative might well understand and respect the customer's concerns but be unable to respond to them because they go against the company's policy. In such cases, a different morality is required, namely one that focuses on the duties and constraints of a role occupant. The sales representative, as an employee of the company, has to be loyal and comply with its policies and procedures. So, from this moral point of view, the aim is not anymore to mediate between conflicting *individual* interests but to meet the interest of a *social* entity, i.e., the group to which one belongs. On the one hand, individual interests are merged into a group interest; on the other hand, this group interest stands against the interests of outsiders. This explains why in a trust game with in-group and out-group differentiation, trust towards members of the in-group is much higher than towards those of the out-group and why a lack of identification with one's group creates suspicion and even harsh discrimination (see Fehrler & Kosfeld, 2013).

The moral currency of Stage 3 is "reputation", which goes beyond inter-individual relationships and spreads throughout the respective social units. An employee, a club member, or a member of some other community may have a certain reputation as a role

14 In this "envelope game", which is a repeated asymmetric game, player 1 receives an envelope in the first stage of each round, containing information on the magnitude of an incentive to defect (high or low). In the second stage, player 2 can decide to continue or end the game. Player 1 has four strategies: (1) cooperate without looking, (2) cooperate with looking, (3) look and cooperate only when the incentive to defect is low, and (4) always defect. "Cooperate without looking" is the strategy by which player 1 can implement the Golden Rule and signal trustworthiness.

occupant. With respect to work, it can relate to the entire organisation or to smaller units like departments or teams. In these contexts, one can win or lose reputation, and in the ultimate case of non-compliance one can be laid off or excluded otherwise.[15]

Stage 3B concerns the interaction between groups or representatives of different groups. Here we focus on broader public life beyond individual group membership, where rules of decency and generalised expectations are concerned. For instance, in many countries there are different rules about how much tip one ought to give in restaurants or on other occasions. When foreigners visit a country, they try to work out what is decent in this respect (see also Azar, 2004). Buchanan's (1994) stress on work ethics and the importance of also preaching them concerns this point because he focuses on generalised cultural rules and expectations (on this see also Rose 2018).

Cultural rules that constitute Stage 3B can be called "conventions", because they are not formally enforced and are rules in some communities but not in others. However, they are not to be confused with "conventions" in the game-theoretic sense (where they apply to coordination games). Stage 3B-rules are conventions in the sense that they can be operative in one community but not in another. In some cultures, a handshake (really) indicates the conclusion of a contract. In other cultures, perhaps no one would rely on it. However, apart from the conventional aspect, there is also the normative aspect concerning the generalised expectations about what a decent person ought and ought not to do (see also Bicchieri, 2006, pp. 39–42; Tomasello, 2016, pp. 126–128). This twofold nature of Stage 3B principles has led to some confusion concerning the question of whether there can be any sharp distinction between social conventions and moral norms (see Sugden, 2010; Brennan, 2010), but as the present discussion reveals, these qualms can be accommodated.[16]

An interesting example in this context relates to "corporate social responsibility". Some fifty years ago or so, companies were much less expected to comply with CSR standards, and it was widely agreed that they had to act in the best interest of the shareholders (Friedman 1970). Today, however, our general expectations have changed in a long process of public discourse about CSR issues, so that the reputational risk for companies as well as possible reputational gains are much higher for companies today compared to the past.

---

15 A good example indicating this is a public goods game where institution formation is possible. Kosfeld, Okada and Riedl (2009) have found that only the grand organisation (where all players participate in the institution) is actually implemented, even though smaller organisations would still be profitable for the participants. Thus, the players forego a monetary benefit, but at the same time solve the so-called second-order free-rider problem (that those not participating in the institution could free-ride on it and earn an even higher payoff than those who implement it). Second-order free-riding can be interpreted in terms of non-compliance with one's role under the institution which applies to the whole group.

16 See also section 6 with respect to distinguishing morality in the economic sense (the middle column in Tab. 2) from morality in the ethical sense (the right-hand column in Tab. 2). Moral norms in the sense of solutions to mixed-motive games belong to the realm of instrumental or strategic reasoning, whereas moral judgments in the ethical sense belong to what is referred to as normative or ethical reasoning.

Another example of the importance of generalised expectations can be seen in the punishment schemes observed in public goods games. Herrmann, Thoeni, and Gächter (2008) as well as Gächter and Herrmann (2011) have found out that cultures differ a great deal in the prevalence of prosocial and antisocial punishment. While the former is efficient, the latter obstructs coordination, because antisocial punishment redeems and crowds out prosocial punishment. Grieco, Faillo, and Zarri (2017) have shown that the most efficient rule would be to allow only the individual with the highest contribution to punish others. However, such a rule would have to be socially shared and sustained by the players, or in other words, become part of their culture.

Finally, whenever people have conflicting views on what is decent or not (perhaps with respect to the punishment regimes just mentioned), the conflict must be resolved by some higher authority constituted by the law or a supreme ruler. Legislators make rules on whether and to what extent one may or may not cover one's face with a veil, may or may not play loud music in residential areas, and so on. In modern societies, we have legislators and judges to solve these disputes. In antiquity, there were kings, queens or other supreme rulers. In large corporations, which also function like smaller societies, the top management is the authority that sets the general rules which all members of the organisation have to follow (e.g., anti-corruption and compliance rules).

Abiding by the rules increases one's reputation; violating them reduces it. However, a failure to comply with these rules may not only spoil an agent's reputation, but in the case of a serious or repeated violation, legal punishment may apply, so that one has to pay a fine or even go to prison. Actual legal punishment can be understood as moving one stage down the moral hierarchy, where the punishment is (or ought to be) regarded as a *decent* response to the offence.

7. Ethics and economics – a taxonomy for the social sciences

If it is true that moral principles function as institutions, we end up with an economic theory of morality that internalises morality into economics as solution concepts for mixed-motive games. However, would this entail that ethics becomes part and parcel of economics? My answer is "no",[17] for the role of moral principles as institutions constitutes but one aspect of morality, and there is yet another, distinctly ethical one, which is *not* captured by the economic approach.

This concerns the idea of justice, pertaining to the normative evaluation of a social situation from an impartial third-person point of view (and to how such an impartial

17 Vanderschraaf (2019; see also Hankins & Vanderschraaf, 2021) raises the same question and answers it in a similar way, however only with respect to the relevance of theories of justice (see below).

point of view can be attained in the first place). For instance, morality in terms of "concerns with avoiding harm, benefitting others, issues of physical violence, emotional hurt, exploitation, subjugation, unequal treatment, unfairness, social injustices, upholding rights, and ... issues of prejudice and discrimination, peace, war, genocide, enslavement, and much more" (Turiel, 2014, p. 4) obviously focuses on the *values* that a moral person should have internalised. In a similar way, Brennan et al. (2013) hold – against Bicchieri (see 2006, pp. 28–35) – that moral norms require such normative attitudes that cause agents to follow a certain norm.

There are clearly (at least) these two distinct meanings of morality, and the difference seems vitally important. *Moral principles as institutions* are tools to solve social dilemmas. They belong to the category of instrumental reasoning (even though from an objective, trans-individual point of view).

As opposed to this, *moral principles as ideas of justice* are "normative" in the ethical sense. It is common in economics to differentiate between positive and normative theory, but as it appears, we have to make a further distinction within this broadly "normative" domain: between the domain of instrumental reasoning, on the one hand, and the domain of ethical reasoning, on the other hand. In order to capture this, now threefold, distinction between domains of reasoning, I refer to

(1) "positive" (or explanatory) reasoning,
(2) "instrumental" (or strategic) reasoning, and
(3) "normative" (or ethical) reasoning.

Normative reasoning in this sense is about answers to questions concerning justice, but also the "good life" in general. The right-hand column of Tab. 2 reveals that such questions can be asked with respect to a single individual, a group of individuals, or with respect to anybody who is affected by a certain decision or state of affairs.

Table 2: A taxonomy for the social sciences (in particular economics and ethics). The fields mentioned in brackets are non-exhaustive examples for the categories.

Theoretical Perspective (Regulative Principle):	Positive/Explanatory (Truth)	Instrumental/Strategic (Effectiveness)	Normative/Ethical (Good Life)
Social Perspective:			
Intra-individual (Decision Theory)	Positive DT (Behav. econ.)	DT (Indiv. rat. choice)	Normative DT (Values, virtues)
Inter-individual (Game theory)	Positive GT (Behav. econ.)	Instrumental GT (Non-coop. GT)	Normative GT (Cooperative GT)
Trans-individual (Soc. systems theory)	Positive SST (Macro econ.)	Instrumental SST (Institutional econ.)	Normative SST (Theories of justice)

Tab. 2 presents a systematic differentiation of research areas in two dimensions. The *first* dimension pertains to regulative principles, i.e., "truth" with respect to explanatory

questions (positive), "effectiveness" with respect to strategic questions (instrumental), and the "good life" with respect to what preferences we ought to have (normative).

The *second* dimension pertains to the social perspective. The *intra-individual* perspective concerns decision theory (DT) asks how we have to explain individual behaviour (positive DT), what an individual with certain preferences should do under specific restrictions (instrumental DT), or what preferences an individual should have in the first place (normative DT). The *inter-individual* perspective concerns game theory (GT) and the question of how to explain the players' choices in games (positive GT), what strategy profiles are rationalisable in games in terms of Nash equilibria (instrumental GT), and what is the best outcome for coalitions of players in terms of what is known as "cooperative game theory" (normative GT). Finally, the *trans-individual* perspective pertains to the analysis of smaller or larger social systems and how they function (positive social systems theory; SST) according to the governing rules of the social entity in question. Instrumental SST is meant to solve cooperation problems by turning mixed motive games into coordination games, i.e., establishing and implementing institutions to realise win-win options. It is one thing to determine what kind of institution is rational to implement from the point of view of the involved agents (instrumental SST), but it is another to determine what is ethically just from an impartial outside point of view (normative SST). For instance, any win-win solution will do as a solution concept from the point of view of instrumental SST, but not all may pass muster from the point of view of normative SST. That latter is what we typically understand as "ethics" in a narrow sense of theories of justice.

In traditional philosophy of science, especially in Karl Popper's "logic of scientific discovery" (1959), positive theory was considered as the only true "science". However, developments in recent years have revealed two important problems with respect to this view: On the one hand, positive theory is not entirely value-free, because we cannot determine how to assign truth-values to theories or statements based on deduction alone (see e.g., Putnam, 2002; see also Minnameier, 2004; 2017; 2023). On the other hand, it has become clear that instrumental and normative questions are scientific questions in their own right. In particular, the different interpretations of rational choice theory (typically labelled "positive" and "normative") relate to structurally different research questions, of which each has its own scientific dignity (Putnam, 2015; see also Hands, 2012; Scanlon, 2014). The present taxonomy tries to capture this, and at the same time it carves out the systematic differences between research areas in which economists, psychologists, philosophers, and social scientists in general are active.

Moreover, as in the title I speak of the "fabric" of social sciences, the taxonomy does not merely reveal structurally different areas of research but also how they are linked to one another. This seems to be crucial with respect to the systematic connections between different fields of research in general, but also with respect to the role(s) of institutions in particular. Let me highlight two systematically different routes that lead to considering and establishing institutions, but for two different reasons and from two distinctively different starting points:

(1) The first route is a move from *normative* issues of justice to *instrumental* reasoning on how we can bring this aspect of justice to bear in real life. In other words, here we move from *normativity* (in the sense of theories of justice) to *implementation* (in the sense of institutional economics).
(2) The second route remains within the *instrumental* domain of strategic reasoning, where players in a game realise that they are caught in a social trap and move from instrumental GT to instrumental SST (as explained in sections 4 and 5).

8. Conclusions and ramifications

The basic idea offered in the present paper is that moral principles function as solution concepts for mixed-motive games and how these games are transformed into coordination games with the help of rather simple moral principles. It has also been shown that these games form a hierarchical order so that each morality-based coordination game ultimately leads to a new mixed-motive game at a higher order. In sections 5 and 6, three elementary moral stages and their substages have been described.

This *strategic-instrumental* account of morality raises the questions of whether and to what extent morality is effectively internalised into economics and what becomes of the classical distinction between *moral* and *instrumental* reasoning. In section 7, a framework for the systematic differentiation of research areas has been presented in which this classificatory problem has been solved. Moreover, this framework allows us to analyse the systematic connections between these research areas and the interdisciplinary transitions from one field to another (as they take place in theory but also in real life). Based on this, we can also solve a few riddles that have been mentioned in the introduction and which have been around for some time in the literature. I would like to address them now.

For instance, we wonder why people, who are certainly all endowed with some sense of morality, are susceptible to moral hypocrisy (Batson et al., 1999; Lönnqvist, Irlenbusch & Walkowitz, 2014; Rustichini & Villeval, 2014; see also Schier, Ockenfels & Hofmann, 2016). The puzzle seems to derive from a confusion of instrumental and normative questions. From a *normative* point of view, it is (perhaps) clear that flipping a coin is a fair way to determine who should be assigned a favourable or an unfavourable task. However, for this to become a rule of the game, it would have to be implemented as an institution, which in this case is prevented because the sanctions cannot be played. Hence, the transfer from *normative* reason to *instrumental* rationality is blocked.

By the same token, we can analyse the issue of the "moral wiggle room". "Plausible deniability" as implemented, e.g., by Dana, Weber and Kuang (2007) or Andreoni and Bernheim (2009) in the dictator game, prevents the receiver from holding the dictator accountable for a zero transfer and thus destroys the higher-order moral game (those who still transfer money do it for strictly *normative* reasons). Accordingly, the sharing norm (Stage 1B) cannot function as an institution under such circumstances..

Understanding moral principles as institutions allows us not only to explain but also to *rationalise* seemingly fickle behaviour. A basic clue for disentangling real-life "(im)moral" agency lies in identifying the proper realms of ethics and economics.

References

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica, 77*, 1607–1636.

Aoki, M. (2011). Institutions as cognitive media between strategic interactions and individual beliefs. *Journal of Economic Behavior and Organization, 79*, 20–34.

Aumann, R. J. (1974). Subjectivity and correlation in randomizing strategies. *Journal of Mathematical Economics, 1*, 67–96.

Aumann, R. J. (1987). Correlated equilibrium and an expression of Bayesian rationality. *Econometrica, 55*, 1–18.

Azar, O. H. (2004). What sustains social norms and how they evolve? The case of tipping. *Journal of Economic Behavior & Organization 54*, 49–64.

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics, 1*, 122–133.

Batson, C. D, Thomson, E. R., Seuferling, G., Whitney H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77*, 525–537.

Beck, K., Dransfeld, A., Minnameier, G., & Wuttke, E. (2002). Autonomy in heterogeneity? Development of moral judgement behaviour during business education. In K. Beck (ed.), *Teaching-learning processes in vocational education: Foundations of modern training programmes* (pp. 87–119). Frankfurt a.M.: Lang.

Becker, G. S. (1976). *The economic approach to human behavior.* Chicago, IL: University of Chicago Press.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge, UK: Cambridge University Press.

Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms*. New York: Oxford University Press.

Binmore, K. (1994). *Game theory and the social contract, vol. 1: Playing fair*. Cambridge, MA: MIT Press.

Binmore, K. (2010a). Social norms or social preferences*. Mind and Society, 9*, 137–159.

Binmore, K. (2010b). Game theory and institutions. *Journal of Comparative Economics, 38*, 245–252.

Binmore, K. (2011). *Natural justice*. Oxford: Oxford University Press.

Bonowski, T., & Minnameier, G. (2022). Trust in impersonal relationships. *Journal of Economic Psychology, 90*, 102513.

Bowles, S. (2016). *The moral economy: Why good incentives are no substitute for good citizens.* Yale University Press.

Brennan, G., & Buchanan, J. M. (1985). *The reason of rules: Constitutional political economy*. Cambridge, MA: Cambridge University Press.

Brennan, G., Eriksson, L. Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. Oxford: Oxford University Press.

Bruni, L., & Sugden, R. (2007). The road not taken: How psychology was removed from economics, and how it might be brought back. *Economic Journal, 117*, 146–173.

Buchanan, J. M. (1964). What should economists do? *Southern Economic Journal, 30*, 213–222.

Buchanan, J. M. (1977). *Freedom in constitutional contract: Perspectives of a political economist.* College Station: Texas A&M University Press.

Buchanan, J. M. (1994). Ethics and economic progress. Norman, OK: University of Oklahoma Press.

Cappelletti, D., Goth, W., & Ploner, M. (2011). Being of two minds: Ultimatum offers under cognitive constraints. *Journal of Economic Psychology, 32*, 940–950.

Carden, A., Caskey, G. W., & Kessler, Z.B. (2022). Going far by going together: James M. Buchanan's economics of shared ethics. Business Ethics Quarterly, 32(3), 359–373.

Casal, S., Falluchi, F., & Quercia, S. (2019). *Journal of Economic Psychology, 70,* 67–79.
Colby, A., & Kohlberg, L. (1987). *The measurement of moral judgment, Vol. 1: Theoretical foundations and research validation.* Cambridge, MA: Cambridge University Press.
Croson, R. (2005). The method of experimental economics. *International Negotiation, 10*, 131–148.
Damon, W. (1977). *The social world of the child.* San Francisco, CA: Jossey-Bass.
Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*, 67–80.
Darwall, S. (2006). *The second-person standpoint: Morality, respect, and accountability*. Cambridge, MA: Harvard University Press.
Deci, E. L., & Ryan, R. M. (2017). *Self-determination theory: Basic psychological needs in motivation, development and wellness.* New York: The Guilford Press.
DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin, 139*, 477–496.
Dietrich, F., & List, C. (2013a). A reason-based theory of rational choice. *Noûs, 47*, 104–134.
Dietrich, F., & List, C. (2013b). Where do preferences come from? *International Journal of Game Theory, 42*, 613–637.
Dietrich, F., & List, C. (2017). What matters and how it matters: A choice-theoretic representation of moral theories. *The Philosophical Review, 126,* 421–479.
Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism: Experimental evidence and new theories. In S. Kolm & J. Ythier (Eds.), *Handbook on the economics of giving, reciprocity, and altruism, Vol. I* (pp. 615—669), Amsterdam: Elsevier.
Fehrler, S., & Kosfeld, M. (2013). Can you trust the good guys? Trust within and between groups with different missions. *Economics Letters, 121*, 400–404.
Friedman, M. (1970). The social responsibility of business is to increase its profits. *New York Times,* September 13, 122–126.
Fudenberg, D, & Tirole, J. (1991). *Game theory*. Cambridge, MA: The MIT Press.
Gächter, S., & Herrmann, B. (2011). The limits of self-governance when cooperators get punished: Experimental evidence from urban and rural Russia. *European Economic Review, 5,* 193–210.
Gert, B. (2005). *Morality: Its nature and justification* (revised ed.). New York: Oxford University Press.
Gintis, H. (2014). *The bounds of reason: Game theory and the unification of the behavioral sciences* (revised ed.). Princeton, NJ: Princeton University Press.
Gneezy, U., & Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies, 29*, 1–17.
Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them.* New York: Penguin.
Greif, A., & Kingston, C. (2011). Institutions: rules or equilibria? In N. Schofield & G. Caballero (eds), *Political economy of institutions, democracy and voting* (pp. 13–43). Berlin: Springer.
Grieco, D., Faillo, M., & Zarri, L. (2017). Enforcing cooperation in public goods games: Is one punisher enough? *Journal of Economic Psychology, 61*, 55–73.
Guala, F., & Hindriks, F. (2015). A unified social ontology. *Philosophical Quarterly, 65*, 177–201.
Guala, F. (2016). *Understanding institutions: The science and philosophy of living together*. Princeton, NJ: Princeton University Press.
Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*, 998–1002.
Hamann, K., Bender, J., & Tomasello, M. (2014). Meritocratic sharing is based on collaboration in 3-year-olds. *Developmental Psychology, 50*, 121–128.
Hands, D. W. (2012). The positive-normative dichotomy and economics. In U. Mäki (Ed.), *Philosophy of economics* (pp. 219–239). Amsterdam: Elsevier.
Hankins, K., & Vanderschraaf, P. (2021). Game theory and ethics. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), URL = <https://plato.stanford.edu/archives/win2021/entries/game-ethics/>.
Herrmann, B., Thoeni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, 319,* 1362–1367.

Hobbes, T. (2001/1651). *Leviathan.* South Bend, IN: Infomotions.
Hodgson, G. M. (2019). Evolution and moral motivation in economics. In M. D. White (ed.), *The Oxford Handbook of Ethics and Economics* (pp. 117–137). Oxford: Oxford University Press.
Hoffman, E., McCabe, K., & Smith, V. (1996). Social distance and other-regarding behavior. *American Economic Review, 86*, 653–660.
Hoffman, M., Yoeli, E., & Nowak, M.A. (2015). Cooperate without looking: Why we care about what people think and not just what they do. *PNAS, 112*, 1727–1732.
Hollis, M., & Sugden, R. (1993). Rationality in action. *Mind, 102*, 1–35.
Homann, K., & Blome-Drees, F. (1992). *Wirtschafts- und Unternehmensethik*. Göttingen: Vandenhoeck & Ruprecht.
Homann, K. & Pies, I. (1994). Wirtschaftsethik in der Moderne: Zur ökonomischen Theorie der Moral. *Ethik und Sozialwissenschaften 5*(1), 3–12.
Homann, K., & Suchanek, A. (2005). *Ökonomik: Eine Einführung*. 2nd ed., Tübingen: Mohr Siebeck.
Johnson, N. D., & Mislin, A.A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*, 865–889.
Kamenica, E. (2012). Behavioral economics and psychology of incentives. *Annual Review of Economics, 4*, 427–452.
Killen, M., Elenbaas, L., & Rizzo, M. T. (2018). Young children's ability to recognize and challenge unfair treatment of others in group contexts. *Human Development, 61*, 281–296.
Kohlberg, L., Levine, C., & Hewer, A. (1983). *Moral stages: A current formulation and response to critics*. Basel: Karger.
Kohlberg, L. (1984). *Essays on moral development, Vol. 2: The psychology of moral development*. San Francisco, CA: Harper & Row.
Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review, 99*, 1335–1355.
Krebs, D. L., & Denton, K. (2005). Toward a more pragmatic approach to morality: A critical evaluation of Kohlberg's model. *Psychological Review, 112*, 629–649.
Lapsley, D. K. (2006). Moral stage theory. In M. Killen & J. G. Smetana (Eds.), *Handbook of Moral Development* (pp. 37–66). New York: Psychology Press.
Ledyard, J. O. (1995). Public goods: A survey of experimental research. In J. H. Kagel & A. E. Roth (Eds.), *Handbook of experimental economics* (pp. 111–194). Princeton, NJ: Princeton University.
List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy, 115*, 482–494.
Lönnqvist, J.-E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: Impression management or self-deception? *Journal of Experimental Social Psychology, 55*, 53–62.
Lütge, C. (2005). Economic ethics, business ethics and the idea of mutual advantages. *Business Ethics: A European Review, 14*(2), 108–118.
Lütge, C. (2012). *Wirtschaftsethik ohne Illusionen*. Tübingen: Mohr Siebeck.
Lütge, C., & Mukerji, N. (eds.) (2016). *Order ethics: An ethical framework for the social market economy.* Springer.
Lütge, C., Uhl. M. (2005). *Wirtschaftsethik*. München: Vahlen.
Meier, S. (2007). A survey of economic theories and field evidence on pro-social behavior. In B. S. Frey & A. Stutzer (eds.), Economics and psychology: A promising new cross-disciplinary field (pp. 51–87), Cambridge, MA: MIT Press.
Minnameier, G. (2000). *Strukturgenese moralischen Denkens – Eine Rekonstruktion der Piagetschen Entwicklungslogik und ihre moraltheoretischen Folgen*. Münster: Waxmann.
Minnameier, G. (2001). A new stairway to moral heaven: A systematic reconstruction of stages of moral thinking based on a Piagetian 'logic' of cognitive development. *Journal of Moral Education, 30*, 317–337.
Minnameier, G. (2012). A cognitive approach to the 'happy victimiser'. *Journal of Moral Education, 41*, 491–508.
Minnameier, G. (2013a). Deontic and responsibility judgments: An inferential analysis. In F. Oser, K. Heinrichs & T. Lovat (Eds.), *Handbook of moral motivation: Theories, models, applications* (pp. 69–82). Rotterdam: Sense.

Minnameier, G. (2013b). Der homo oeconomicus als „happy victimizer". *Zeitschrift für Wirtschafts- und Unternehmensethik, 14*, 136–156.
Minnameier, G. (2014). Moral aspects of professions and professional practice. In S. Billet, C. Harteis & H. Gruber (Eds.) *International handbook of research in professional and practice-based learning* (pp. 57–77). Berlin: Springer.
Minnameier, G. (2016). Rationalität und Moralität – Zum systematischen Ort der Moral im Kontext von Präferenzen und Restriktionen. *Zeitschrift für Wirtschafts- und Unternehmensethik, 17*(2), 259–285.
Minnameier, G. (2017). Forms of abduction and an inferential taxonomy. In L. Magnani & T. Bertolotti (Eds.) *Springer handbook of model-based reasoning* (pp. 175–195). Berlin: Springer.
Minnameier, G. (2020). Explaining happy victimizing in adulthood – A cognitive and economic approach. *Frontline Learning Research, 8*(5), 70–91.
Minnameier, G. (2023). The logical process and validity of abduction. In L. Magnani (ed.), *Handbook of abductive cognition* (pp. 159–180). Springer Nature.
Minnameier, G. (2025). Ordonomik und ökonomische Theorie der Moral: Eine kritische Verhältnisbestimmung im pädagogischen Kontext. In ders. (Hrsg.), *Ordonomik als Beitrag zur Bildung für gesellschaftliche Verantwortung* (S. 175–200). Bielefeld: wbv.
Minnameier, G., Beck, K., Heinrichs, K., & Parche-Kawik, K. (1999). Homogeneity of Moral Judgement? Apprentices solving business conflicts. Journal of Moral Education, 28, 429–443.
Mulder, L. B. (2009). The two-fold influence of sanctions on moral norms. In D. De Cremer (Ed.), *Psychological perspectives on ethical behavior and decision making* (pp. 169–180). Charlotte, NC: Information Age Publishing.
Mulder, L. B., van Dijk, E., De Cremer, D., & Wilke, H. A. M. (2006). Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas. *Journal of Experimental Social Psychology, 42*, 147–162
Mulder, L. B., van Dijk, E., Wilke, H. A. M., & De Cremer, D. (2005). The effect of feedback on support for a sanctioning system in a social dilemma: The difference between installing and maintaining the sanction. *Journal of Economic Psychology, 26*, 443–458.
Paulus, M., & Moore, C. (2014). The development of recipient-dependent sharing behaviour and sharing expectations in preschool children. *Developmental Psychology, 50*, 914–921.
Paulus, M., & Moore, C. (2017). Preschoolers' generosity increases with understanding of the affective benefits of sharing. *Developmental Science, 20*, e12417.
Pies, I. (2000). *Ordnungspolitik in der Demokratie: Ein ökonomischer Ansatz diskursiver Politikberatung.* Tübingen: Mohr Siebeck.
Pies, I. (2009). *Moral als Heuristik: Ordonomische Schriften zur Wirtschaftsethik*. Berlin: wvb.
Pies, I. (2016). The ordonomic approach to order ethics. In C. Lütge & N. Mukerji (eds.).
Pies, I. (2022). *30 Jahre Wirtschafts- und Unternehmensethik: Ordonomik im Dialog*. Berlin: wvb.
Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson, London.
Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays*. Cambridge, MA: Harvard University Press.
Putnam, H. (2015). Naturalism, realism, and normativity. *Journal of the American Philosophical Association, 1*, 312–328.
Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118*, 57–75.
Rand, D. G., Greene, J. D., & Nowak, M.A. (2012). Spontaneous giving and calculated greed. *Nature, 489*(7416), 427–430.
Rose, D. C. (2018). *Why culture matters most.* New York, NY: Oxford University Press.
Rustichini, A., & Villeval, M. C. (2014). Moral hypocrisy, power, and social preferences. *Journal of Economic Behavior and Organization, 107*, 10–24.
Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica, 15,* 243–253.
Scanlon, T. M. (2014). *Being realistic about reasons*. Oxford: Oxford University Press.

Schelling, T. C. (1960). *The strategy of conflict.* London: Oxford University Press.
Schier, U. K., Ockenfels, A., & Hofmann, W. (2016). Moral values and increasing stakes in a dictator game. *Journal of Economic Psychology*, *56*, 107–115.
Smith, Craig E, & Warneken, Felix (2016). Children's reasoning about distributive and retributive justice across development. *Developmental Psychology*, *52*(4), 613–628.
Sugden, R. (2005). *The economics of rights, co-operation and welfare* (2nd ed.). Houndmills: Palgrave Macmillan.
Sugden, R. (2010). Is there a distinction between morality and convention? In M. Baurmann, G. Brennan, R. E. Goodin & N. Southwood (Eds.), *Norms and values: The role of social norms as instruments of value realization* (pp. 47–65). Baden-Baden: Nomos.
Sugden, R. (2018). *The community of advantage: A behavioural economist's defence of the market.* Oxford: Oxford University Press.
Tomasello, M. (2016). *A natural history of human morality*. Cambridge, MA: Harvard University Press.
Turiel, E. (2014). Morality: epistemology, development, and social opposition. In M. Killen & J. G. Smetana (Eds.), *Handbook of moral development* (pp. 3–22). Mahwah, NJ: Lawrence Erlbaum Associates.
Ullmann-Margalit, E. (1977). *The Emergence of Norms*. Oxford: Clarendon Press.
Vanberg, V. (1986/1994). Spontaneous market order and social rules: A critical examination of F. A. Hayek's theory of cultural evolution. In V. Vanberg, Rules and choice in economics (pp. 77–94). London: Routledge.
Vanberg, V. (1988/1994). Morality and economics: De moribus est disputandum. In V. Vanberg, Rules and choice in economics (pp. 41–59). London: Routledge.
Vanberg, V. (2008). On the economics of moral preferences. *American Journal of Economics and Sociology*, *67*(4), 605–628.
Vanberg, V., & Buchanan, J. M. (1988/1994). Rational choice and moral order. In V. Vanberg, Rules and choice in economics (pp. 60–73). London: Routledge.
Vanderschraaf, P. (1995). Convention as correlated equilibrium. *Erkenntnis*, *42*, 65–87.
Vanderschraaf, P. (2019). *Strategic justice: Convention and problems of balancing divergent interests.* New York: Oxford University Press.