Synthese
https://doi.org/10.1007/s11229-021-03333-y

ORIGINAL RESEARCH

Rationality in games and institutions

Philippe van Basshuysen[1]

Received: 30 September 2020 / Accepted: 23 July 2021

Abstract
Against the orthodox view of the Nash equilibrium as "the embodiment of the idea that economic agents are rational" (Aumann, 1985, p 43), some theorists have proposed 'non-classical' concepts of rationality in games, arguing that rational agents should be capable of improving upon inefficient equilibrium outcomes. This paper considers some implications of these proposals for economic theory, by focusing on institutional design. I argue that revisionist concepts of rationality conflict with the constraint that institutions should be designed to be incentive-compatible, that is, that they should implement social goals in equilibrium. To resolve this conflict, proponents of revisionist concepts face a choice between three options: (1) reject incentive compatibility as a general constraint, (2) deny that individuals interacting through the designed institutions are rational, or (3) accept that their concepts do not cover institutional design. I critically discuss these options and I argue that a more inclusive concept of rationality, e.g. the one provided by Robert Sugden's version of team reasoning, holds the most promise for the non-classical project, yielding a novel argument for incentive compatibility as a general constraint.

Keywords Rationality · Concept formation · Team reasoning · Game theory · Institutional design · Robert Sugden

1 Introduction

How do rational agents interact with each other? This question has occupied economists and philosophers alike, and game theory has become a common language in which their answers are formulated. Theorists adhering to 'classical' concepts of rationality in games answer this question by asserting that rational strategies in a

This article belongs to the topical collection "Concept Formation in the Natural and Social Sciences: Empirical and Normative Aspects", edited by Georg Brun, Catherine Herfeld, and Kevin Reuter.

✉ Philippe van Basshuysen
philippe.v.basshuysen@gmail.com

[1] Institute of Philosophy, Leibniz University Hannover, Hannover, Germany

Published online: 18 October 2021

game are those that support equilibria as defined by Nash (1950), or refinements thereof (e.g. Harsanyi & Selten, 1988). Unfortunately, in many games there are only inefficient equilibria, as the notorious Prisoners' Dilemma reminds us. There is a growing literature proposing that the concept of rationality in games be reformed, based on the argument that this concept should allow for the possibility that rational agents can bring about outcomes that improve upon inefficient equilibria. These 'non-classical' concepts are underpinned by theories of practical rationality, such as constrained maximization (Gauthier, 1987) or team reasoning (e.g. Bacharach, 2006) (both of which will be introduced in the next section).

This paper considers some implications for economic theory if the proposed non-classical concepts were to be widely adopted, by focusing on the practically important field of institutional design. More precisely, I shall draw attention to a widely accepted constraint on institutional design: wherever possible, institutions should be designed to be *incentive-compatible*, that is, the incentives of the agents interacting through these institutions should be aligned with the social goals that the institution is designed to bring about; or, stated negatively, it should not create adverse incentives that may bring about unintended social consequences. Game theory plays a central role in the methodology of incentive-compatible institutional design: designers devise games that implement social goals in equilibrium, and they aim to make institutions resemble those games (Guala, 2001). This methodology is based on a classical concept of rationality, and on a rationality assumption: it presumes that rational agents follow equilibrium strategies, and that the people who will interact through the designed institution are rational. Together, these assumptions imply that the designed institution will bring about the desired social goals.

What would the adoption of a non-classical concept of rationality imply for institutional design? Through an extended example, I show that many non-classical theories—in particular those that are put forward as rivals to classical rationality rather than as complements to it—are at odds with the constraint of incentive compatibility. That is, when imposing this constraint, institutional designers' expectations concerning individual behavior are in many cases inconsistent with the recommendations of non-classical rationality.

I then identify three strategies that proponents of strong conceptions of non-classical rationality could pursue to resolve this conflict: (1) reject incentive compatibility as a general constraint on institutional design; (2) contend that the individuals interacting through the designed institutions are irrational; or (3) concede that their revisionist concepts do not cover institutional design. I draw out the negative consequences of each option. First, rejecting the constraint of incentive compatibility would risk diminishing social welfare, which is a concern that is supported empirically and by principled arguments. Second, treating people as irrational would violate a core principle of economic theory, namely the assumption that people are instrumentally rational (Herfeld, 2020), which most proponents of re-engineered rationality concepts are unwilling to abandon. Third, conceding that non-classical theories do not cover institutional design would seem to be a serious blow to non-classical concepts. Concerning the third option, the positive contribution of this paper is to provide an interpretation of some moderate, non-classical theories—such as Robert Sugden's theory of team reasoning—which could block the conclusion

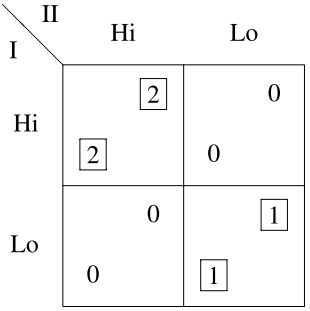


Fig. 1 Example of a Hi-Lo game. Player I's payoff is shown on the bottom left and player II's payoff on the top right of each cell. A square around a payoff number denotes a player's best reply to the opponent's choice

that non-classical concepts do not cover institutional design. When these theories are interpreted as providing an inclusive concept of rationality, reflecting the possibility that both classical and non-classical behavior can be rational, they are in line with treating incentive compatibility as a general constraint and may provide a novel argument for doing so.

The modest lesson I draw from this analysis is that revisionist programs concerning concepts that occupy important places in social scientific theories should be evaluated in relation to what changes they would necessitate in those theories and in how they bear on the social world, and whether these changes would be desirable. Starkly non-classical projects do not seem to pass this test in relation to institutional design.

The next section presents classical and non-classical concepts of rationality in games and Section 3 introduces incentive-compatible institutional design. Section 4 brings together the previous two sections: it shows that non-classical concepts of rationality conflict with incentive compatibility, and it discusses the ways in which proponents of revisionist concepts might resolve this conflict, arguing that a more inclusive account of rationality might be the preferable strategy. Section 5 concludes by urging some caution concerning starkly non-classical projects.

2 Concepts of rationality in games

Concepts of rationality in games derive from theories that prescribe how games should rationally be solved. Classical theories place the Nash equilibrium, or refinements thereof, at center stage, while proponents of revisionist programs ground their proposed reconceptualization in alternate theories, which share the implication that rationality allows agents to improve upon inefficient equilibria. I will first present classical theories of rationality in games, followed by constrained maximization and team reasoning as two paradigmatic representatives of non-classical theories.

In order to introduce these theories, it will be helpful to consider some simple games in normal form. Figure 1 shows the "Hi-Lo game"; two agents—players I and II—simultaneously choose Hi or Lo. Each combination of their choices results in a cell, specifying both players' payoffs; player I's payoff is shown on the bottom left and player II's payoff on the top right of each cell. The payoff numbers refer

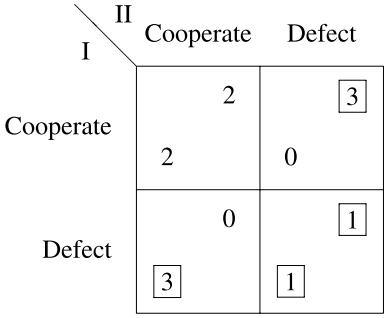
to the players' utilities, which are measured on an interval scale, that is, the point of zero utility and the units of measurement are arbitrary, and the ratios of the differences between utilities are non-arbitrary. Interpersonal comparability of players' utilities need not be assumed. Furthermore, the utilities are not transferable between the players. What should player I choose? This apparently depends on what player II chooses: if II chooses Hi, I's best reply is to play Hi too, but if II chooses Lo, I's best reply is to play Lo; equivalently for player II. In Fig. 1, a player's best replies to her opponent's possible choices are marked with squares around her payoff numbers. There are two outcomes in the game in which the players' actions are best replies to each other: both play Hi, or both play Lo. Furthermore, the players may also come up with strategies in which both available actions are chosen with non-zero probabilities; for example, they could choose their actions depending on the outcome of the throw of dice. If players' strategies include such plans of action in which they mix their choices, there is a third outcome in which the players' strategies are best responses to each other: both players play Hi with probability 1/3 and Lo with probability 2/3, which yields both players an expected payoff of 2/3.

In general, a strategy profile in which each player's strategy is a best reply to all the other players' strategies (where strategies may be pure or mixed), is a *Nash equilibrium* (or simply "equilibrium" in the following). Call individual strategies that are played with a positive probability in some equilibrium, "equilibrium strategies". According to classical theories of rationality, then, a strategy is rational only if it is an equilibrium strategy. Because in equilibrium, no player has incentives to deviate from her strategy, classical theories establish a tight link between rationality and incentives. This leads Robert Aumann to state the classical view thus:

> The Nash equilibrium is the embodiment of the idea that economic agents are rational; that they simultaneously act to maximize their utility. If there is any idea that can be considered *the* driving force of economic theory, that is it. Thus in a sense, Nash equilibrium embodies the most important and fundamental idea of economics, that people act in accordance with their incentives. (Aumann, 1985, p. 43, emphasis in original)

According to the most common classical theory, equilibrium strategies are not only necessary, but also sufficient for rationality (e.g. Binmore, 2007). This theory is sometimes contested because games can have many equilibria, some of which may be better than others for all players. For instance, in the Hi-Lo game, despite there being three equilibria, it seems "trivial" to many that (Hi, Hi) should be the unique rational outcome of this game (e.g. Gold & Sugden, 2007, p. 284), and "paradoxical" that standard game theory does not solve for this outcome alone (e.g. Bacharach, 2006, p. 44 et seq.). It should be noted, however, that even though this critique sometimes motivates a departure from classical theories (e.g. Gold & Sugden, 2007, p. 284 et seq.), refinements of Nash equilibrium, according to which playing equilibrium strategies is necessary, but not sufficient for rationality, can rule out "bad" equilibria. John Harsanyi and Reinhard Selten, for example, developed a general theory that selects a unique equilibrium in a large class of games (including all games in normal form) (Harsanyi & Selten, 1988). In Hi-Lo games, their concept of

Fig. 2 Example of a Prisoners' Dilemma game



"payoff dominance" implies that only the Hi strategy is rational to play, thus ruling out the two inefficient equilibria.

In contrast to classical theories, according to which equilibrium strategies are a necessary condition for rational play (and in some of which equilibrium strategies are also a sufficient condition for rational play), non-classical theories maintain that equilibrium strategies are neither necessary nor sufficient for rational play. These theories are motivated by a rejection of the implication that rationality results in inefficiency in some games. These theories thus imply that rational players will improve on these outcomes, at least under certain conditions. Hi-Lo games might not necessitate the adoption of this type of theory (because some classical theories can accommodate this concern),[1] but Prisoners' Dilemma games bring out this alleged shortcoming of classical theories. An example of a Prisoners' Dilemma is shown in Fig. 2. In this game, it is a dominant strategy for both players to defect, thus (Defect, Defect) is the unique equilibrium. But (Cooperate, Cooperate) strictly Pareto-dominates this equilibrium, that is, if the players were to achieve this outcome, both would be better off. According to classical theories, rational players could never achieve this outcome, since both would have incentives to deviate by defecting. Instead of interpreting these incentives as a constraint on what can rationally be achieved, proponents of non-classical theories consider the fact that rational players will reach inefficient equilibria to be a weakness of the orthodoxy. Let us now look at two non-classical theories, constrained maximization and team reasoning, in more depth.

2.1 Rationality-as-constrained-maximization

David Gauthier defends the view that rational players can improve upon inefficient equilibria in situations where there is grounds for mutual trust. He locates rationality at the level of agents' dispositions to choose (rather than at the level of

[1] However, Harsanyi and Selten's theory does not provide much of a justification of why rational players will reach equilibria that payoff-dominate others. Some non-classical theories can be interpreted as providing these justifications and might thus be used to justify particular refinements. I thank an anonymous reviewer for pointing this out.

strategies for choices), and he defines "constrained maximization" as the disposition to choose cooperatively if the other agent(s) have the identical disposition, and non-cooperatively otherwise.[2] A population of constrained maximizers playing Prisoners' Dilemmas could thus improve upon the non-cooperative outcome that "straight maximizers" achieve, who simply choose best replies.

This view faces the problem that straight maximizers could simply exploit constrained maximizers in Prisoners' Dilemma games. The latter would then be worse off as a consequence of their disposition to cooperate, which would constitute an odd account of practical rationality. Gauthier excludes this possibility by introducing the condition of "translucency". This means that an individual's disposition toward constrained or straight maximization is known to others in the population with a positive probability. Thus, when individuals play against each other, there is a probability that constrained maximizers will recognize each other and cooperate, and there is a probability that a constrained maximizer will fail to recognize a straight maximizer and will therefore be exploited. According to Gauthier, it is rational for individuals to choose the disposition to constrain their maximization if this maximizes their expected utility. This is the case, roughly, when the choice of disposition is sufficiently translucent and the proportion of constrained maximizers in the population sufficiently large.[3]

In a nutshell, constrained maximization seeks to rationalize cooperation by placing rationality at the level of dispositions and requiring that these dispositions be translucent. The argument is supposed to rationalize non-equilibrium play in Prisoners' Dilemmas and, more generally, in all games in which an outcome strictly Pareto-dominates all equilibria. Moreover, in games, such as Hi-Lo, in which there is a Pareto efficient equilibrium, the theory solves for this equilibrium. In short, for Gauthier, Pareto efficiency takes the place of equilibrium as the criterion for rationality.

2.2 Rationality-as-team-reasoning

Furthermore, a number of theorists have criticized concepts of rationality that derive from best-reply reasoning as too individualistic. These theorists argue that players may sometimes reason from the perspective of "we", instead of "I", that is, as a team.[4] When they reason as a team, players identify the action profiles (instead of individual actions) that best promote the common interests of the team

[2] The canonical exposition of his theory of constrained maximization is in chapter 6 of Gauthier (1987). Recent attempts at rationalizing cooperation in Prisoners' Dilemma games are Gauthier (2013, 2015). I critically discuss the latter analysis in van Basshuysen (2017).

[3] For constrained maximization to maximize utility requires a combination of the two conditions—level of translucency and proportion of constrained maximizers in the population—such that the more constrained maximizers there are, the higher the risk can be that they fail to recognize straight maximizers. For a detailed exposition, see Gauthier (1987, p. 176 et seq.).

[4] The following are some of the main contributions to team reasoning in games. Robert Sugden introduced team reasoning to game theory in Sugden (1993). Michael Bacharach developed the theory formally in Bacharach (1999). Bacharach (2006), which was completed by Natalie Gold and Sugden after Bacharach passed away, can be seen as the culmination of Bacharach's theory, connecting team reasoning to findings in social psychology. Hurley (2005a, b) developed a theory of team reasoning according

they form part of; they then choose the individual actions that jointly generate those action profiles. For instance, if the players of a Prisoners' Dilemma game form a team, they may identify mutual cooperation as their preferred action profile and choose to cooperate in order to bring it about. Team reasoning presupposes that individuals identify with the team that they jointly constitute, which has also been described as a transformation from individual to collective agency (Gold & Sugden, 2007, p. 292). As a result of this transformation, the players put aside their individual interests and act upon the interests of the team. This process raises two questions: how, exactly, do individual interests convert into team interests? And why would rational players act upon the team interests, which might (depending on the answer to the first question) require them to sacrifice individual utility to benefit the team?

On the first question, Robert Sugden (2011, 2015) argues that team play should yield the players a *mutual advantage*, requiring that the outcome is at least as good as the players' maximin payoff (that is, the utility that each player can achieve independently of the other players). A possible formal characterization of mutual advantage has been provided by Karpus and Radzvilas (2018). They present a measure that can be applied to calculate which outcome(s) of a normal form game maximize the players' mutual advantage, relative to possible reference points. By reaching action profiles that maximize mutual advantage, or some other measure of team interests, rational players would implement outcomes, which, in many games, yield higher utility to the players than equilibrium outcomes. But this would often require them to choose contrary to their incentives, for instance, when cooperation is identified as the outcome that best advances the team interest in a Prisoners' Dilemma game. Why would rational players do this? According to proponents of team reasoning, as a consequence of this "agency transformation", the joint action of the team can be described as rational; entailing that the individual choices of the team members, which constitute this joint action, are rational too. There is disagreement amongst proponents about whether agency transformation is itself a requirement of rationality. For instance, Susan Hurley (2005a, b) defends a strong version of team reasoning, contending that team identification is itself the result of rational choice. Others, such as Sugden (2003) and Michael Bacharach (2006), argue that team identification or the failure thereof is not a matter of rational choice, but comes prior to it—as there would be no objective of choice without a unit of agency. Their theories can be interpreted as supporting a more inclusive concept of rationality because, when the units of agency are "singletons" (that is, individuals do not identify as a team with other individuals), what is rational to choose, according to these theories, coincides with what classical rationality would prescribe—while this may not be the case when individuals identify as (non-singleton) teams. For reasons that will be spelled out below, I am more optimistic about non-classical theories yielding inclusive rationality concepts than those that exclude classical strategies altogether; for

Footnote 4 (continued)
to which the unit of agency is itself a matter of rational choice. For comparisons of different theories of team reasoning, see Gold and Sugden (2007) and Karpus and Gold (2017). I present what I take to be the core tenets of team reasoning here, rather than some details over which debate continues.

now, however, it suffices to note that, according to all theories of team reasoning, individuals can identify as teams, in which case it may be rational to play non-equilibrium strategies.[5]

To sum up this section, we have distinguished classical concepts of rationality from non-classical concepts, by introducing orthodox solution concepts in game theory and contrasting them with rationality-as-constrained-maximization and rationality-as-team-reasoning.[6] The distinguishing feature is that classical concepts bind rational choice to equilibrium strategies, whereas the theories supporting non-classical concepts seek to rationalize non-equilibrium strategies that may lead to outcomes that Pareto dominate equilibria. Even though non-classical theories arrive at their conclusions for different reasons—team reasoning through group identification, constrained maximization through individuals' disposition to choose—they share some of those conclusions. For instance, both imply that there are conditions under which rational agents cooperate in Prisoners' Dilemma games. Let's next introduce the second cornerstone of the argument, namely, institutional design.

3 Institutional design and the constraint of incentive compatibility

Institutions can develop "spontaneously", but they can also be designed towards specific social goals, in particular, generating social welfare. In the latter case, a designer seeks to devise an institution to govern some interactions in a way that will, if successful, maximize welfare. Typically, designers model the interactions in question as a game, using game theory to determine how the interactions would result in different outcomes depending on the rules of the game, or *mechanisms*, and then choose the one that best promotes the defined goals. Designers strive to make mechanisms *incentive-compatible* in order to reliably produce desirable outcomes; let's consider an example from Roger Myerson (2008):

The seller of an indivisible item faces one potential buyer. It is commonly known that the seller values the item either at \$0 or at \$80, and that the buyer values the item either at \$100 or at \$20. Let's assume that the traders simply seek to maximize their profits, that is, dollar values define their utilities. The traders both profit from trade unless a type 80 seller (a "strong seller") faces a type 20 buyer (a "strong buyer"). Whether they are weak or strong is their private information, but it is

[5] It should be added that team reasoning is sometimes put forward as a descriptive theory of interactive choice, rather than as a theory of rationality. For instance, when Sugden states his own position in Bacharach (2006), he emphasises that he is "less concerned with the validity of team reasoning, treating it only as an idealised model of a form of reasoning which people in fact use, whether justifiably or not" (xxii). However, in other places he suggests that according to team reasoning, "the rationality of each individual's action derives from the rationality of the joint action of the team" (Sugden, 2003, p. 167; also cf. Gold & Sugden, 2007, p. 285). Sugden might thus be interpreted as conceiving of team reasoning as a theory of rationality, at least in some contexts. Be that as it may, I am interested here in team reasoning insofar as it is put forward as a theory of rationality, as in Hurley's account.

[6] These are not the only non-classical theories that have been proposed, but they may be the ones most widely discussed in the literature.

	Buyer's value	
	[strong]	[weak]
Seller's value	$20	$100
[strong] $80	0, *	1, $90
[weak] $0	1, $10	1, $50

P(trade), E(price if trade)

Fig. 3 Split-the-Difference mechanism. From Myerson (2008)

commonly known that each is of a strong or a weak type with an independent probability of 1/2.

The social planner wishes to come up with a mechanism that determines, depending on the traders' valuations, whether trade should happen, and if so, at what price. Since the valuations are the traders' private information, a mediator will ask them for their valuations. Depending on the information provided, she will announce whether and at what price the item will be sold. What could this mechanism look like? An obvious idea is the "split-the-difference mechanism"; whenever the buyer's valuation is higher than the seller's, the item will be sold for a price at the midpoint between the two valuations, and if both are strong (i.e. the buyer's valuation is smaller than the seller's), the item won't be sold. The table in Fig. 3 shows this mechanism. Each cell corresponds to a combination of the players' types. The first number in a cell denotes the probability that the item will be sold, and the second number the price in cases in which it is sold.

But will the traders report their types honestly? The standard methodology of institutional design proceeds on the assumption that this cannot be expected because, for instance, a weak seller (i.e. one that values the item at $0) would gain from lying about her type: assuming that the buyer is honest, the seller's expected profit from revealing weakness is $1/2(10-0)+1/2(50-0)=30$; if she instead claims to be strong, her expected profit is $1/2(90-0)=45$ (and the same argument applies for the weak buyer). Thus, it is not an equilibrium of split-the-difference that the traders honestly reveal their types: this mechanism is incentive-*in*compatible.

Conversely, a mechanism that, unlike split-the-difference, does make it an equilibrium for players to reveal their types is incentive-compatible. Continuing the example, we can ask what constraints an incentive-compatible mechanism must satisfy. For simplicity, let's make the following assumptions, which are shown in the table in Fig. 4. As before, if the seller and the buyer are both strong, the probability of trade is 0, and if both are weak, the item will be sold for $50 with probability 1. If one is weak and the other strong, the trade will occur with a probability q that does not depend on who is weak or strong. If the trade occurs, the profit of a weak trader

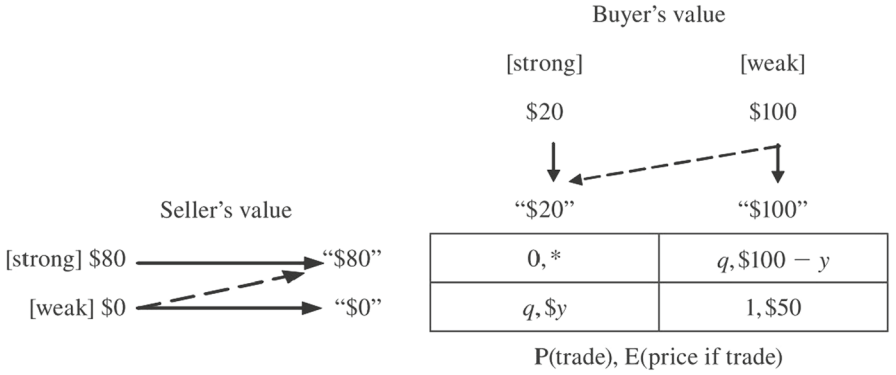


Fig. 4 Symmetric scheme with parameters q and y. From Myerson (2008)

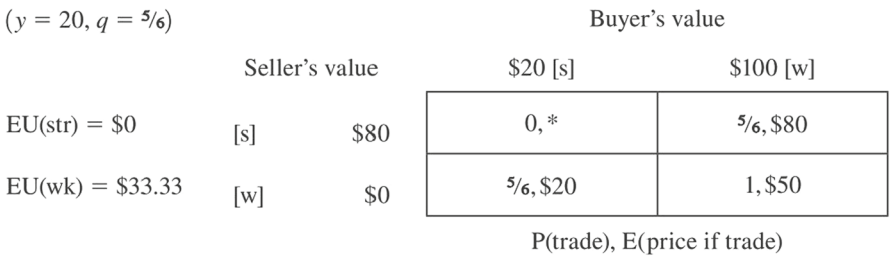


Fig. 5 The 5/6-mechanism, which is incentive-compatible. From Myerson (2008)

against a strong trader is some number y, which again is the same no matter who is weak or strong, as can be seen in the lower-left and the upper-right cell of the table.[7]

With these assumptions in place, we can ask what constraints the parameters q and y must satisfy to make the mechanism incentive-compatible. First, note that a strong buyer would only buy the item for a price smaller than (or equal to, suppose) $20. Similarly, a strong seller would sell the item only for a price larger than or equal to $80. Therefore, the parameter y must satisfy the participation constraint $y \leq 20$.

For honesty to constitute an equilibrium, we must make it an optimal response for traders to honestly reveal their types if they expect the other trader to honestly reveal their type too. It can be verified in Fig. 4 that a strong seller or buyer would never gain by claiming to be weak. But, depending on the parameters y and q, a weak seller or buyer might gain by claiming to be strong. Consider the weak buyer. Her expected payoff from honesty is $1/2(q)(y) + 1/2(50)$, and her expected payoff from lying is $1/2(q)(100 - y)$. Thus, for honesty to be an equilibrium, the parameters q and

[7] The buyer and the seller are thus treated symmetrically. This assumption is only for simplicity; see Myerson (2008, p. 595) for how it can be relaxed.

y must satisfy the incentive constraint $1/2(q)(y) + 1/2(50) \geq 1/2(q)(100 - y)$, which reduces to $q \leq 25/(50 - y)$.[8]

Thus, a mechanism that satisfies the participation constraint $y \leq 20$ and the incentive constraint $q \leq 25/(50 - y)$ makes honest participation an equilibrium and is therefore incentive-compatible. Setting $y = 20$ achieves the largest feasible probability for the trade to happen, viz. $q = 5/6$. This mechanism, call it the 5/6-mechanism, is shown in Fig. 5. The expected profits (ex ante, that is, before the types are revealed) for each trader in this mechanism are $[0 + 0 + (5/6)20 + (1)50]/4 = 16.67$.

This example shows how incentive compatibility delimits the feasible amount of social welfare: the 5/6-mechanism yields a positive probability that the item will not be traded when the traders are of different types, even though this means that the item will not go to the trader who values it most. So ex post, after the traders reveal their types, the mechanism produces allocative inefficiencies in cases in which the traders are of different types but the trade does not occur. We saw that there is no incentive-compatible mechanism with a lower probability of such allocative inefficiencies than the 5/6-mechanism.[9] Thus, this mechanism determines the boundary of feasible social welfare when agents are strategic.

The same point can be made by comparing expected profits. There is no incentive-compatible mechanism that would give both traders a higher expected profit than the \$16.67 that the 5/6-mechanism yields: in technical terms, this mechanism is *ex ante incentive efficient*. If it were possible to rely on the players' honesty and dispense with incentive compatibility, then the split-the-difference mechanism would yield an expected profit of \$17.5 for both players ($[0 + 10 + 10 + 50]/4 = 17.5$), thus improving upon the incentive efficient mechanism. However, institutional designers do not expect individuals who will interact through the mechanism to reveal their private information honestly unless they are provided incentives to do so. For this reason, they seek to design the institutions that govern social interactions to be incentive compatible whenever possible; even if doing so delimits the social welfare that can be generated through the institutions.[10]

Note that, while the problem of incentive compatibility has been introduced here with regard to agents revealing their private information, the problem is generally faced by social planners. Of every institution whose outcomes depend on individuals revealing private information or performing actions that the social planner cannot fully observe, it can be asked whether the institution gives those individuals incentives to reveal their information, or to perform their hidden actions obediently. Since it is hard to think of an institution that does not rely on individuals' private information and in which all the individuals' actions can always be fully observed, the problem of incentives is ubiquitous in institutional design.

[8] It can easily be verified that this constraint is identical for the seller.

[9] In fact, the "revelation principle" implies that, if the traders are strategic, there is no mechanism (incentive-compatible or not) in which the probability of allocative inefficiency is lower than in the 5/6 mechanism.

[10] The limits to social welfare can be examined more generally: see Holmström and Myerson (1983).

4 Non-classical concepts conflict with incentive-compatible institutional design

As we bring together non-classical concepts of rationality and incentive-compatible institutional design, we arrive at an apparent conflict: there are cases in which what institutional designers expect individuals to do is inconsistent with what they would do if they were "non-classically" rational. In the example from the previous section, the constraint of incentive compatibility implies that the expected utilities of the two traders equal 16.7 at best (that is, in the 5/6-mechanism); but if the traders were to reason as a team, or to constrain their maximization, they could commit to revealing their types in the split-the-difference mechanism and they could reap the benefits from this commitment, receiving a mutually advantageous expected utility of 17.5. If the designer's goal is to maximize the traders' expected utility, she would thus only implement an incentive-compatible mechanism if she were not expecting the traders to be honest when they can profit from lying in split-the-difference. More generally, when institutional designers treat incentive compatibility as a general constraint, they do not believe that the individuals interacting through the institutions follow team reasoning or constrained maximization: for, if they did, they would expect that more social welfare could be achieved when dropping the constraint. The following three claims thus form an inconsistent triad: (i) institutions should be designed so as to maximize social welfare; (ii) incentive compatibility should be required as a general constraint on institutional design; and (iii) the individuals interacting through the designed institutions are non-classically rational.

Thus, in order to preclude inconsistency, (i), (ii) or (iii) must be rejected. I take it for granted that rejecting (i), that institutions should be designed so as to maximize social welfare, should be excluded as an ethically unreasonable position. If this is correct, we are left with two options: reject the assumption that institutions should be designed to be incentive-compatible (ii), or that individuals interacting through the designed institutions are non-classically rational (iii). Institutional designers would typically reject (iii) since, as we have seen, they treat incentive compatibility as a general constraint, basing their methodology on a classical rationality assumption. What about proponents of a reformed rationality concept? They might either reject (ii), that incentive compatibility should be required as a general constraint on institutional design; or, they might take exception to the claim that individuals interacting through the designed institutions are non-classically rational (iii). There are two ways in which (iii) could be rejected because one could negate either the "rational"-part, or the "non-classical"-part of the claim. That is, non-classical theorists could claim that the individuals interacting through the designed institutions are *irrational*, where the meaning of "rational" is fixed by their non-classical theory. Or they could assert that these individuals are rational, but not in a non-classical meaning of the word; in which case they would seem to concede that their theories do not apply when it comes to individuals interacting through institutions.

In summary, in order to avoid inconsistency, and assuming that we should design institutions to maximize social welfare, proponents of a reformed rationality concept must make one of the following claims:

(1) incentive compatibility should not be required as a general constraint on institutional design;
(2) the individuals interacting through the designed institutions are irrational; or
(3) institutional design should be covered by a classical rationality concept.

Let's consider these options in turn.

4.1 Away with incentive compatibility?

Some reformers contend that incentive compatibility should not function as a general constraint on institutional design. Thus, Sugden, commenting on this paper, argues that "mechanisms should be incentive-compatible with respect to the agents that exist (i.e. mechanisms should provide individual incentives to individuals, team incentives to teams). So I can't see why a [team reasoning] theorist can't happily take the first option and reject individual incentive compatibility in cases in which individuals identify as team members." This is a well-motivated proposal since, as we have seen, (individually) incentive-compatible mechanisms would sacrifice social welfare compared to some incentive-*in*compatible mechanisms if individuals were to identify as a team and to act on the associated team interests; and thus, if we knew that they will form a team in a given case, we had better implement the mechanism that is (individually) incentive-*in*compatible but incentive-compatible with respect to the team of players.

This proposal faces an epistemic challenge, however, because designers do not know in advance whether those interacting through the institution to be designed, will identify as individuals or as teams.[11] Arguably, they should then base the decision of whether to retain or let go of incentive compatibility on empirical evidence (cf. Hausman, 1998): should individuals be expected to behave in line with classical rationality in the institution to be designed, perhaps because in similar settings they have been found to behave in this way? Or is there evidence that in some kinds of institutional settings, they systematically identify as teams and that we could base our design on this assumption? Unfortunately, there is a paucity of research on this latter question; there appear to be no empirical studies to date on institutional or mechanism design on the assumption that team reasoning or constrained maximization should yield the rationality standards. In contrast, whether individual behavior approaches classical rationality when individuals interact through institutions has been studied extensively. In a survey of these studies, Daniel McFadden (2009) shows that individual behavior meets the expectations of classical rationality in institutional settings when incentives are large, even when the choice structure is complex. In contrast, when incentives are small and ambiguous, deviations from these expectations grow, which McFadden attributes to individuals' putting less effort into determining best replies and being more distracted by irrelevant factors.

[11] Below, I will argue that Sugden's theory could also be interpreted in a different vein, thus avoiding this epistemic challenge.

These findings suggest that classical rationality may be a good approximation in institutional settings and thus, that incentive compatibility is important, especially where the stakes are high for the individuals who interact through the institution in question, whereas there might be room for relaxing this constraint in some lower-stakes contexts.

While indicating that classical rationality may be a good approximation in many institutional settings, these findings fall short of establishing incentive compatibility as a general constraint, rather, this might be thought to call for more research on human behavior in the context of institutional design. Since it is less-than-certain that individual behavior will approach classical rationality in the institution to be designed, would it not appear that the standard methodology—assuming classical rationality and treating incentive compatibility as a general constraint—faces a similar empirical challenge to Sugden's proposal?

However, the epistemic challenge is less severe with respect to standard methodology because there is a principled argument to be made for the importance of incentive compatibility, as is shown by the fact that even some pessimists about the empirical adequacy of rational choice theory have made a case for incentive compatibility. Alexander Rosenberg (1992), for instance, argues that economic theory is predictively weak but nevertheless normatively valuable, in designing the institutions through which we interact. According to Rosenberg, what matters for design purposes are not actual institutional outcomes but counterfactual outcomes if everyone were classically rational, because institutions must be robust: they must work even in the case in which everyone were, in Hume's words, a "knave" (1742). Geoffrey Brennan and James M. Buchanan (1983; 1985) make similar arguments. The basis of these arguments is a kind of precautionary reasoning[12]: as we have seen, incentive compatibility delimits feasible social welfare; but incentive-*in*compatible institutions produce adverse incentives that may diminish welfare *much more*. This point can be made more precise in the example from the previous section. We saw that the incentive-compatible, 5/6-mechanism yields both players an expected profit of 16.67, which is the maximum feasible welfare when the traders are strategic. While players could reach a higher expected profit (17.5) in the incentive-*in*compatible, split-the-difference mechanism, if they could commit to being honest about their types, their loss in welfare would be considerably more severe if they failed to act on their commitment: there is an equilibrium in which both traders falsely report strong types when they are weak with a probability of 3/5, and the expected payoff in this equilibrium is only 10 for both traders.

The asymmetry of the possible welfare losses provides a reason for treating incentive compatibility as a general constraint where actual behavior is uncertain; or, in other words, a social planner should require incentive compatibility as a

[12] Hausman (1998) criticizes these arguments, maintaining that the normative value of economic theory for the design of institutions should only depend on the empirical question of whether the theory accurately predicts behavior. We need not take a stance here on this disagreement: since Hausman presents himself as more optimistic about the predictive accuracy of orthodox rational choice theory than Rosenberg, or Brennan and Buchanan, this position might imply that institutions should be designed to be incentive-compatible.

default, which should be overridden only if there is *very good evidence* that individual behavior will approach the recommendations of non-classical rationality in particular institutional arrangements and that we can thus go without it. So the burden of proof appears to be on proponents of non-classical concepts who contend that we should go without it. Furthermore, this point may reinforce the empirical evidence for the importance of incentive compatibility, as surveyed by McFadden (2009), suggesting, at the very least, that classical rationality is a good approximation in institutional settings and that non-classical approaches will thus have a hard time living up to their burden of proof. In combination, these arguments go some way towards urging that the first option, that is, doing away with incentive compatibility, should be resisted. While this result is contrary to Sugden's proposal, I will suggest below that Sugden's own theory of team reasoning also lends itself to a different interpretation, which would support treating incentive compatibility as a general constraint.

4.2 Institutions for the irrational?

If a proponent of non-classical rationality is reluctant to reject incentive compatibility as a general constraint, they could reason thus: "we should not expect individuals interacting through institutions to follow the ideals of non-classical rationality; rather, we should expect them to follow best-reply reasoning, and for this reason, incentive compatibility should generally be required. Because the individuals cannot reap the fruits of cooperation, they are *irrational*." This is the second possible option for aligning non-classical rationality with incentive compatibility as a general constraint.

This imagined proponent of non-classical rationality, however, is likely to be a fiction; for one of the fundamental principles of economic theory, which is almost universally adopted, is the assumption that individuals are instrumentally rational (Herfeld, 2020), and reformers of the rationality concept are unlikely to be willing to give up this assumption, often stressing that human beings can, and often do, meet their rationality standards. For instance, Gauthier endorses the "conception of human beings as rational (or potentially rational) individual actors" (1987, p. 93). In fact, qualms with the classical assumption that rational individuals cannot achieve optimal outcomes in certain games, such as Prisoners' Dilemmas, appear to be among the main reasons for proposing non-classical concepts in the first place. Proponents of these concepts demand that stringent rationality requirements be applied, and it would be an odd view to demand this while denying that people can meet these requirements when interacting through institutions. The option of denying people's rationality thus looks rather dismal.

However, perhaps this is premature, as it might be possible to spell out this option more favorably, at least for some versions of non-classical theories. So far, I have described non-classical theories as implying that agents are irrational whenever they do not constrain their maximization, or do not reason as a team. While this is true for some versions of these theories (e.g. Hurley might be seen as a proponent of this view), for other versions, the assumption would suffice that only some, not

all, individuals are irrational, in order for their theory to be consistent with incentive compatibility as a general constraint. For example, remember that constrained maximizers will rationally defect in Prisoner's Dilemmas when a sufficiently large fraction of the population are defectors, and depending on the level of translucency. Similarly, if a fraction of agents renege on their commitment to reveal information honestly or to act obediently in an incentive-*in*compatible institution, it might be rational for the others to do the same, according to constrained maximization. For some varieties of team reasoning, similar arguments could be constructed. For instance, according to Sugden (2015), in order to follow team reasoning, a player requires assurance that the other player(s) do so as well—which might not be the case if other players are irrational. Anticipating that individuals would rationally defect in an incentive-*in*compatible institution in the presence of irrational individuals, proponents of these theories could argue that the institution should be designed to be incentive-compatible. This line of argument nevertheless commits proponents of non-classical concepts to the assumption that *some* of the individuals are irrational, which might be deemed undesirable. Alternatively, these proponents could retain incentive compatibility for the same reason that drivers wear seatbelts: we don't expect the accident to happen, but why take a chance? Similarly, imposing incentive compatibility would be the consequence of a kind of precautionary reasoning: "while we don't expect this, there *might* be a decisive fraction of irrational individuals who won't cooperate, which may result in universal defection, even when other people are rational. In the face of this uncertainty, we had better take preventive action and impose incentive compatibility as a general constraint, because failing to do so entails the risk of bringing about socially undesirable outcomes."

While some proponents of non-classical concepts might adopt this strategy in order to bring their theories in line with incentive-compatible institutional design, it comes at a cost. There is an asymmetry between classical theories of rationality and non-classical theories when combined with this strategy; while under the former, incentive compatibility should be imposed as a constraint *because* people are assumed to be rational, under the latter this would be seen a constraint because people are potentially irrational and *despite* the expectation that they are rational. Thus, while under classical theories, the rationality principle provides the rationale for treating incentive compatibility as a constraint, this is not the case for non-classical theories; in these theories, the rationale for this constraint would rather be provided by the possibility that individuals are irrational, that is, by the negation of the rationality principle. For theorists committed to this principle, this puts non-classical theories at a disadvantage. Furthermore, given the significance of the rationality principle for economic theory, and the fact that it is almost universally adopted, this strategy is unlikely to attract many followers among reformers of the rationality concept.

4.3 (Halfway) back to the orthodoxy

If a proponent of a revisionist concept of rationality does not want to reject incentive compatibility as a general constraint (claim (1.) above), or to assume that individuals

interacting through the designed institutions are irrational (claim (2.)), this seems to be a self-defeating position, at least at first glance: they want incentive compatibility and they want to treat people as rational, but their rationality standards don't allow for both. Their remaining option seems to be to overthrow their standards, that is, to concede that rational individuals should not be expected to constrain their maximization or to reason as a team when interacting through institutions. In other words, they would concur that non-classical rationality does not cover institutional design; rather, a concept of classical rationality should apply here. They would seem to overthrow their project of reforming the rationality concept and become proponents of the orthodoxy.

However, it is possible to provide an alternative interpretation of some (though not all) non-classical theories, which would block this stark conclusion. We could interpret these theories as allowing for a more inclusive concept of rationality; attempting to rationalize some courses of action which classical theories deem outright irrational, without denying the rationality of classical behavior.[13] Indeed, some of the non-classical theories that we encountered earlier can be interpreted in a way that accommodates such an inclusive concept.[14] For instance, as we have seen, for Sugden and Bacharach, rational choice is conditional on the choice of agency, and when the units of agency are singletons, their theories yield classical rationality as a special case. Furthermore, since, in their theories, the choice of agency is not itself subject of rational deliberation, there is nothing irrational about choosing in line with classical rationality; rather, choices are irrational if conditioned on the "wrong" combinations of agency and reasoning (e.g. if a player were to defect in a Prisoners' Dilemma while identifying and seeking to attain the best outcome for a team, or if a player were to cooperate in this game while acting as a singleton). Thus, their theories can be interpreted as entailing that neither classical nor non-classical strategies are outright irrational; that is, as rationalising *more* strategies than the classical conception of rationality allows. But if more strategies, including classical strategies, can be rationalised in this way, there is, as a matter of logic, no conflict with treating incentive compatibility as a general constraint.

While Sugden's theory, in particular, lends itself to this inclusive interpretation, it should be noted that it does not imply Sugden's proposal above, according to which incentive compatibility should be given up in cases where individuals identify as a team. In contrast, I suggest that the inclusive interpretation is not only consistent with treating incentive compatibility as a general constraint, but that it may also provide a novel argument for doing so. If both classical and non-classical strategies are deemed rational and if, in addition, we assume that people will behave rationally in this inclusive sense, how should a mechanism be designed? An important goal may be not to introduce incentives whereby classical and non-classical recommendations

[13] I am much indebted to an anonymous reviewer for suggesting this interpretation and for providing the better part of the arguments to follow.

[14] Other non-classical theories we encountered above, such as Hurley (2005a, b), or Gauthier (2015), do not seem to provide for an inclusive concept as they appear to entail that, when their recommendations diverge from classical behaviour, the latter should be deemed irrational.

come apart; for otherwise the mechanism might fail to align incentives with the social goal in question even when agents are rational. So suppose the designer's goal is to align the incentives of classical and non-classical reasoners in this way; then the designed mechanism should not be incentive-*in*compatible, because if it were, classical reasoners may in some cases "defect" (in our running example, lie about their types when they could profit from it), whereas non-classical reasoners would "cooperate" (be honest about their types), which would contradict the designer's goal that their incentives be aligned. Because the result would thus be equivalent to proceeding on the assumption that people are individually (classically) rational, this may provide a non-classical argument for treating incentive compatibility as a general constraint when non-classical theories are interpreted as sustaining an inclusive concept of rationality.

5 Conclusion

The aim of this paper was to draw out some implications that a re-engineering of the rationality concept would entail in relation to institutional design. I argued that starkly revisionist programs are at odds with the standard methodology of treating incentive compatibility as a general constraint on institutional design. Proponents of these programs have three options to resolve this conflict, all of which seem to be undesirable from their perspective. First, they could reject incentive compatibility as a general constraint, but they would thereby risk wasting social welfare. Second, they could keep this constraint, while treating the individuals interacting through the designed institutions as irrational; yet, this option looks rather unappealing from the perspective of the proponents of non-classical concepts and it shows an advantage of classical concepts, which are in line with the constraint of incentive compatibility. Third, they could grant that their revised rationality concept does not cover institutional design, which would, however, seem to present an argument for classical concepts of rationality. The modest lesson I draw from this analysis is that revisionist programs concerning concepts that occupy important places in social scientific theories should be evaluated relative to what changes they would necessitate in those theories and in how they bear on the social world, and whether these changes would be desirable. Facing three undesirable options, starkly non-classical projects do not seem to pass this test in relation to institutional design. A more modest interpretation of some non-classical accounts, such as Sugden's, would warrant more optimism, allowing for a more inclusive concept of rationality, which is consistent with treating incentive compatibility as a general constraint; indeed, non-classical accounts may then provide a novel argument for this standard methodology.[15]

Funding Open Access funding enabled and organized by Projekt DEAL.

[15] For invaluable discussions and comments, I thank Eric Angner, Luc Bovens, Jurgis Karpus, Donal Khosrowi, Bryan Roberts, Kai Spiekermann, Bob Sugden, Lucie White, and two anonymous reviewers.

References

Aumann, R. J. (1985). What Is game theory trying to accomplish? In K. J. Arrow & S. Honkapohja (Eds.), *Frontiers of economics.* Basil Blackwell.

Bacharach, M. (1999). Interactive team reasoning: A contribution to the theory of co-operation. *Research in Economics, 53*, 117–147.

Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton University Press. Edited by Natalie Gold and Robert Sugden.

Binmore, K. (2007). *Playing for real: A text on game theory*. Oxford University Press.

Brennan, G., & Buchanan, J. (1983). Predictive power and the choice among regimes. *Economic Journal, 93*(369), 89–105.

Brennan, G., & Buchanan, J. (1985). *The reason of rules*. Cambridge University Press.

Gauthier, D. (1987). *Morals by agreement*. Clarendon.

Gauthier, D. (2013). Twenty-Five On. *Ethics, 123*, 601–624.

Gauthier, D. (2015). How I learned to stop worrying and love the prisoner's dilemma. In M. Peterson (Ed.), *The prisoner's dilemma.* Cambridge University Press.

Gold, N., & Sugden, R. (2007). Theories of team agency. In F. Peter & H. B. Schmid (Eds.), *Rationality and commitment.* Oxford University Press.

Guala, F. (2001). Building economic machines: The FCC auctions. *Studies in History and Philosophy of Science Part A, 32*(3), 453–477.

Harsanyi, J. C., & Selten, R. (1988). *A general theory of equilibrium selection in games*. MIT Press.

Hausman, D. M. (1998). Rationality and Knavery. In W. Leinfellner & E. Köhler (Eds.), *Game theory, experience, rationality; Foundations of social sciences; Economics and ethics: In honor of John C. Harsanyi* (pp. 67–79). Kluwer.

Herfeld, C. (2020). Understanding the rationality principle in economics as a functional a priori principle. *Synthese, 198*, 3329–3358.

Holmström, B., & Myerson, R. B. (1983). Efficient and durable decision rules with incomplete information. *Econometrica, 51*(6), 1799–1819.

Hume, D. (1742). Of the independency of parliament. In E. F. Miller (Ed.), *Essays, moral, political and literary.* Liberty Fund, Inc. 1987.

Hurley, S. (2005a). Social heuristics that make us smarter. *Philosophical Psychology, 18*, 585–612.

Hurley, S. (2005b). Rational agency, cooperation and mind-reading. In N. Gold (Ed.), *Teamwork: Multidisciplinary perspectives* (pp. 200–215). Palgrave Macmillan.

Karpus, J., & Gold, N. (2017). Team reasoning: Theory and evidence. In J. Kiverstein (Ed.), *The Routledge handbook of philosophy of the social mind* (pp. 400–417). Abingdon: Routledge Taylor Francis.

Karpus, J., & Radzvilas, M. (2018). Team reasoning and a measure of mutual advantage in games. *Economics & Philosophy, 34*(1), 1–30.

McFadden, D. (2009). The human side of mechanism design: A tribute to Leo Hurwicz and Jean-Jacques Laffont. *Review of Economic Design, 13*, 77–100.

Myerson, R. B. (2008). Perspectives on mechanism design in economic theory. *American Economic Review, 98*(3), 586–603.

Nash, J. F. (1950). Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences, 36*(1), 48–49.

Rosenberg, A. (1992). *Economics—mathematical politics or science of diminishing returns*. University of Chicago Press.
Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy, 10*(1), 69–89.
Sugden, R. (2003). The logic of team reasoning. *Philosophical Explorations, 6*(3), 165–181.
Sugden, R. (2011). Mutual advantage, conventions and team reasoning. *International Review of Economics, 58*, 9–20.
Sugden, R. (2015). Team reasoning and intentional cooperation for mutual benefit. *Journal of Social Ontology, 1*(1), 143–166.
van Basshuysen, P. (2017). The prisoner's dilemma. *Economics and Philosophy, 33*(1), 153–160.