Homo Oeconomicus (2022) 39:127–144
https://doi.org/10.1007/s41412-021-00117-0

RESEARCH PAPER

# Ken Binmore: Behavioral Scientist

**Luciano Andreozzi**[1]

Received: 20 May 2021 / Accepted: 13 December 2021 / Published online: 18 January 2022


**Abstract**
This paper discusses Ken Binmore's contribution to the debate on other-regarding preferences with reference to his contributions on equilibrium selection in non-cooperative games. We first assess his claim that the experimental evidence in favor of different types of social preferences has been vastly exaggerated. Then, we compare Binmore's contribution with some recent developments of the literature. We show that recent experimental evidence lends support to his view that subjects' behavior is mostly driven by a combination of learning and selfishness. From a theoretical point of view, we show that Binmore's positions foreshadowed what is today known as the Social Heuristic Hypothesis.

**Keywords** Other regarding preferences · Game theory · Social heuristics hypothesis

**JEL Classification** D01 · D64 · D83 · D91

## 1 Introduction

Economic models of human interactions have been traditionally built on the three pillars of self-interest, optimality and equilibrium. None of these assumptions holds true, if taken literally. One of the long-standing issues in economic methodology is then if (and why) such unrealistic models are useful in understanding the behavior of actual human beings in their ordinary environment. For more than three decades, Ken Binmore has been a vocal supporter of a radical position on this matter. His view is the combination of two different theses, each of which is quite radical in itself. On the one hand, Binmore upheld an extremely skeptical stance about the power of human rationality to explain social interactions. Binmore has a rather colorful way of expressing his views, that are sometimes worth quoting in full. In a paper that deals with equilibrium selection in Rock-Scissor-Paper kind of games, for example, we read:

✉ Luciano Andreozzi
luciano.andreozzi@unitn.it

[1] Universitá degli Studi di Trento, Trento, Italy

> There has been some debate about the extent to which Von Neumann was anticipated by the great mathematician Emile Borel. This debate is significant here only to the extent that the record shows that Borel formulated the minimax theorem but decided that it was probably false. It therefore seems pointless to run experiments designed to test the hypothesis that laboratory subjects are capable of duplicating Von Neumann's reasoning. (Binmore et al., 2001, p. 445)

So, if game theory is useful at all in explaining human behavior, it is because, given enough time to experiment and learn, average human beings may become better than Emile Borel at playing simple parlour games. However, and this is the crucial caveat, they will learn to *play* better than Borel, without *knowing* anything about calculus or probability, let alone Nash equilibria. Game theory explains ordinary people's optimal choices just like physics explains the way a football player manages to use Newton's laws (together with many other physical facts about, for example, attrition) to score an almost impossible free kick.

This is the second leg of Binmore's approach: a sustained confidence in the power of learning and evolution to explain human behavior in the recurrent situations in which ordinary people interact day-by-day. This position was not new within economics and Binmore was ready to acknowledge his debts with a long series of precursors that trace back to Bernard de Mandeville, Adam Smith and David Hume. In the years in which he put forward his proposal, however, many economists where turning their back to this tradition and Binmore's voice became increasingly isolated. The mounting evidence collected in the then new field of experimental economics seemed to prove that game theory did a very poor job at predicting human behavior, even in simple games. This created the apparent need for a new approach, that would complement (or, for some authors, replace) the discredited "neoclassical economics". Over the years, the heterogeneous approach known as behavioral economics came to dominate the field and Binmore ended up in the rather awkward position of being a defender of the orthodoxy within economics.

The bitter and prolonged debate that ensued contained at least two paradoxes. First, Binmore argued that behavioral economists were committed to the *homo oeconomicus* ideal more than the average neoclassical economist. For example, they clenched to equilibrium refinements such as subgame perfection, that had been discredited by evolutionary game theorist years before. However, and this is the second paradox, behavioral economists were themselves committed to an evolutionary approach to economics, although they disagreed with Binmore on how social evolution was best modelled [see for example Gintis et al. (2005)]. So both camps ended up accusing the other of giving support to a discredited scientific approach (neoclassical economics) that both claimed needed to be replaced by a more rigorous one based on bounded rationality and evolution.

In this paper I reconstruct Binmore's position in this debate, steering clear of the overheated controversies that it generated. The main contention of this paper is that Binmore's many contributions, that are scattered among several technical papers, are best read having in mind the broader picture that emerges from his more philosophical works. I will contend that, taken together, these works propose a coherent

view of how social norms contribute to maintain the co-operation we observe in our societies. Such a view was a sober alternative to the (more relaxed) approach that became dominant in the literature.

This reconstruction is not just an historical curiosity. In the course of exposition I will discuss recent experimental evidence that lends support to Binmore's view that subjects' behavior in social dilemmas is best explained as a mixture of learning and selfishness, rather than in terms of social preferences. I will also show that Binmore anticipated the so called *Social Heuristics Hypothesis*, an approach that has gained increasing attention in the last decade. I shall contend that a careful reading of Binmore's contributions to repeated games would contribute to a more nuanced view on this important topic than the one that emerges from the current literature.

The paper proceeds as follows. In Sect. 2 I will briefly outline Binmore's position concerning the central role reciprocity plays in the maintenance of social order. Section 3 presents his position in the debate about the so-called refinements of Nash equilibria. This discussion sets the stage for a presentation, contained in Sect. 4, of his views concerning the experimental evidence on social preferences. Section 5 compares Binmore's position in this debate with the Social Heuristics Hypothesis. Section 6 concludes.

## 2 It's All About Reciprocity

The most pressing question in all the social sciences, not only in economics, is just what keeps society together. What is the cement that prevents our societies from descending into Hobbesian chaos? (Elster, 1989, p. 1–2). Binmore's answer to this question is contained in a single word: *reciprocity*.[1] Most of the cooperation that we observe among unrelated human beings is explained by the fact that we play similar games over and over with the same people. When this is the case, today's opportunistic behavior triggers negative responses from others tomorrow. Breaches of the prevailing social norms bring about loss of reputation, the break of valuable cooperative relations and even outright punishment. When society works smoothly, it is usually in one's own best interest not to violate the tenet of the prevailing social contract. As usual, Binmore's position is best presented using his own words.

> As a little boy, I can remember wondering why the shopkeeper handed over the candy when I had handed over my money. Why not just keep the money? Hume's answer is that even dishonestly inclined shopkeepers honor the convention that they supply the candy after being payed because they would otherwise risk being punished. In my case, the shopkeeper would have lost more than my custom. I would have told all and sundry about my mistreatment. The damage to his reputation would then be out of all proportion to the small gain in cheating a little boy. (Binmore, 2020, 85)

[1] The word "reciprocity" is used in different ways in the literature. Binmore's variety corresponds to what is usually termed *direct reciprocity* and must be distinguished from *strong reciprocity*, which is more popular among defenders of the social preferences approach. See for example Bowles and Gintis (2013).

Explanations like these are usually dismissed as psychologically implausible. If asked, the shopkeeper would say that handing a little boy the candy after receiving the money is just the right thing to do and his answer would be a perfectly honest one. But in Binmore's view game theoretical models answer a deeper question: why does such a norm of honesty survive? Why do not shopkeepers learn to ignore their moral feelings?[2] To find an answer to this question, one is to make the small exercise of counterfactual thinking that is at the hearth of the notion on Nash equilibrium. What would happen to a shopkeeper who decides to stop handing the candy whenever he receives a payment? Will he have a substantially larger income at the end of the month? This question is important, because it seems plausible that if shopkeepers could become rich by bulling their little customers, this norm of honesty would be eroded over time. Binmore's thesis is based on the intuitively appealing principle that either the gains from breaching a norm are negligible, or that norm is inherently unstable and is doomed to disappear.

It follows that the shopkeeper and the game theorist can be right at the same time. The shopkeeper is right when he says that he deals fairly with his young customers because of a deep-seated sense of justice. The game theorist is right when he remarks that, if the shopkeeper would ever be willing to try, he would discover that treating a boy unfairly is rarely a good idea. After all, very few businessmen got their riches by mistreating their customers.

This example illustrates Binmore's position in the old *homo sociologicus* vs. *homo oeconomics* debate (Binmore & Samuelson, 1994b). According to the common stereotype, believers in homo sociologicus maintain that human beings follow mechanically whatever social norm prevails in the society they live in. As Jon Elster once put it, they prefer to study the fence around the cow rather than the behavior of the cow within the fence (Elster, 1979, 114). In our example, this amounts to take the shopkeeper's answer as a primitive and explain his behavior in term of the social norm of honesty he has been taught to follow. No explanation is given of why the norm we observe has that content and not a different one, nor why we observe norms of this type in certain contexts and not in others.

The game theoretic explanations favored by the supporters of *homo oeconomicus* have a clear advantage here, because they treat social norms as endogenous. Their argument is that the only social norms that we are likely to observe are those that correspond to stable Nash equilibria. However, the homo oeconomicus explanation is obviously wrong if stated in terms of the shopkeeper's intentions and motivations. Who would seriously claim that a shopkeeper performs sophisticated cost-benefit analysis before handing the candy to a boy? Binmore's position is that game theorists are not committed to such a patently absurd view. Game theory explains why social norms remain stable over time, despite the ever present temptation individuals may have to infringe them. To wit, long-run profit maximization explains why the shopkeeper is right in following his norm of honesty without asking too many questions. For this reason

[2] Game theory also explains how norms emerge. I have no space in this note to address this issue.

> one should make no sharp distinction between homo economicus and homo sociologicus. In using a social norm in a situation to which it is well adapted, homo sociologicus behaves as though he were optimizing. Similarly, when optimizing, homo economicus behaves as though he were employing a social norm that is well adapted to his problem. (Binmore & Samuelson, 1994b)

The idea that morality can be reduced to a set of norms that co-ordinate the choices of self-interested individuals on one of the many equilibria of the game of life[3] was not new and it had always met intense resistance. It runs against the more agreeable vision that human beings care (at least to a certain extent) about the public good and the welfare of their fellow citizens. Defenders of this view have been traditionally looked at with suspicion. As we shall see, Binmore would not have been an exception.

## 3 Equilibrium Refinements

An early discovery in game theory was that virtually any game of interest has more than one equilibrium, and most of the times these equilibria are not strict. This happens all the times in which, given the strategy chosen by the opponent, a player has several alternative best replies. If in the game under consideration the players choose sequentially, for example, almost all Nash equilibria will be non strict. As a consequence, a lot of effort has been put into spelling out the conditions that make some equilibria more likely to be observed than others. Most of these efforts had been made in the original tradition of game theory, that emphasizes players' rationality. They produced a plethora of different definitions, some of which, like *subgame perfect* and *sequential* equilibrium, are still in use today.

Since its inception, Binmore had been one of the most vocal dissenters of the literature on *refinements* of Nash equilibrium. This was partly due to the concern for realism he voiced in the passage we quoted at the beginning of this paper. But there was a deeper concern. Despite its mathematical sophistication, the rationalistic approach to game theory was rife with paradoxes and loose ends. The notion of common knowledge was a particularly fertile ground for the production of paradoxes and so was the notion of backward induction (Binmore, 1990).

For this reason, Binmore became one of the main supporters of the evolutionary approach to equilibrium selection, that was gaining traction in the late 80.s. As it is well-known today, this approach had been pioneered by John Nash in his Ph.D. thesis, but it failed to attract attention in the first two decades of research in game theory (Weibull, 1997). The diffusion of evolutionary methods in economics was helped by the fact that in the 70.s biologists had imported game theory in their discipline (Smith, 1982). Now economists were importing back some of the concepts and methods that had been developed within biology in the previous two decades.

The main thrust of the evolutionary approach is that an equilibrium can be considered a serious candidate for selection only if it can be *learned* by individuals

[3] See Bicchieri (2006) for a modern view on social norms based on this approach.

with a limited degree of rationality. In formal terms, this means that credible Nash equilibria need to be stable under an *evolutionary* adjustment dynamics. On this as in many other cases, economists were less fortunate than biologists, because in the social sciences it is harder to give a reasonably clear-cut characterization of what counts as an "evolutionary" dynamics.[4] Loosely speaking, the general agreement in the literature was that to be plausible an evolutionary dynamics should satisfy at least the requirement that strategies that earn a larger payoff should grow over time at the expenses of the strategies that earn smaller payoffs.[5]

The first results in this literature were encouraging. It was not difficult to show, for example, that if an evolutionary adjustment dynamics converges to a state, then that state must be a Nash Equilibrium. Also, under all evolutionary adjustment dynamics strictly dominated strategies where bound to get extinct. However, these results did not extend to the more sophisticated notions of equilibrium employed in the refinements literature. In a series of articles published during the 90.s together with Larry Samuelson (Gale et al., 1995; Binmore & Samuelson, 1994a, 1999) Binmore showed that there is no way to prove, for example, that a population of boundedly rational players will inevitably converge towards a subgame perfect Nash equilibrium. Similarly, evolutionary pressures were insufficient to eliminate weakly dominated strategies. The results for other criteria like *forward induction* was similarly disappointing.

Although in the same years similar results were obtained by other scholars (Cressman & Schlag, 1998), Binmore was the first to see their broader implications for our understanding of the working of social norms and hence of human cooperation at large. In all the articles that he wrote on this topic, Binmore used the Ultimatum (Mini) Game (UG) as the main working-horse model. In the literature, the UG was routinely presented as the prototypical situation in which subjects' behavior in experimental settings could only be explained postulating either irrationality or some non self-interested motives like altruism or inequity aversion. The learning approach Binmore promoted suggested a third, more parsimonious, alternative. Learning models can be technically demanding, but the intuition behind Binmore's result requires little more than simple algebra. Consider a simplified setting in which the UG is played for four euros. Proposers can only make a High (H) offer, in which the four euros are split equally, or a Low (L) offer in which the proposer takes three euros for herself. responders must decide whether to Accept (A) or Reject (R) a Low offer. (High offers are automatically accepted). If a Low offer is rejected, both players receive nothing. This is known as the Ultimatum Mini-Game and its normal form version is represented in Fig. 1 (*Left*).

As it is often the case in laboratory setting, imagine that this game is played anonymously within a (relatively) large population of subjects divided between proposers

[4] "A fundamental point is that biologists almost always deal with the genetic mechanism of natural selection. This mechanism admits a simple, canonical dynamical representation [...] For economists the social mechanisms of learning and imitation are usually more important than the genetic mechanism. A wide variety of learning and imitation processes are conceivable and the appropriate dynamical representation seems to be highly context-dependent." (Friedman 1991).

[5] See Weibull (1997) and Fudenberg and Levine (1998) for early surveys of the literature and Sandholm (2010) for a more recent treatment.

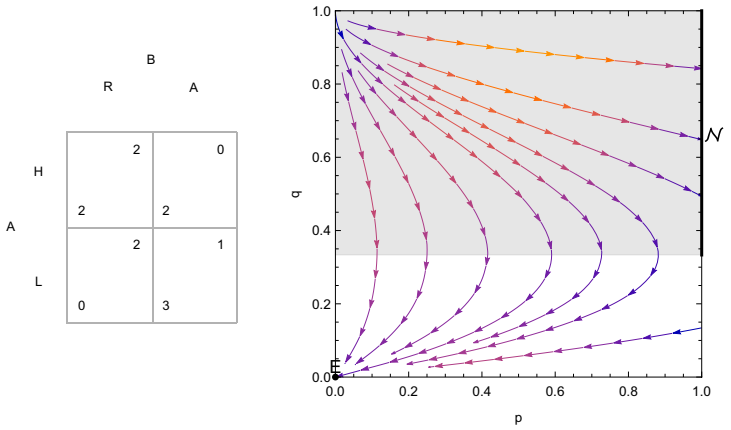


**Fig. 1** (*Left*) The Ultimatum Mini-Game. (*Right*) The Replicator Dynamics may converge to the set $\mathcal{N}$ which is made by non-subgame perfect Nash Equilibria, rather than to the only subgame perfect Nash equilibrium $E$

and responders. Notice that, because interactions are anonymous, players cannot develop a reputation for being willing to reject unfair offers. The only role repetition plays in this setting is to allow subjects to learn which strategies are played by the other players and what are the best responses to them. The *state* of the population is defined by the fraction $q$ of responders who reject unfair offers and the fraction $p$ of proposers who make fair offers. Every possible state of the population corresponds to a point in the square in Fig. 1 (*Right*)). As individuals gain experience with the game, the fractions $p$ and $q$ are subject to change. The arrows in the figure represents the expected evolution of the state of the population under the so-called Replicator Dynamics: the fraction of subjects using each strategy grows at a rate which is proportional to the difference between the payoff of that strategy and the average payoff (Hofbauer & Sigmund, 1998; Weibull, 1997). Hence, those strategies that perform better than average grow at the expenses of the strategies that perform worst than average.

The orbits reveal that the fraction $q$ of responders who Reject unfair offers is always decreasing. This is not surprising: turning down a Low offer is costly, and subjects who play Accept always earn a larger payoff than those who Reject. Instead, the fraction $p$ of proposers who make a High offer grows when Low offers are mostly rejected (that is when $q > \frac{1}{3}$, that is in the shadowed area) while decreases when Low offers are mostly accepted.

Depending on the initial conditions, orbits converge towards two distinct sets of rest points. The first one is point $E = (0, 0)$ in Fig. 1 (*Right*). It corresponds to the subgame perfect Nash equilibrium in which Low offers are made and they are accepted. Other orbits converge to the thick segment $\mathcal{N}$ in the picture, in which all proposers make High offers and at least one-third of the responders reject Low offers. These states correspond to the non-subgame perfect Nash equilibria in which the proposer makes a High offer because she believes that a Low offer would be rejected with a sufficiently large probability.

It is instructive to see what drives this result. Consider a state in the vicinity of the set $\mathcal{N}$. To be concrete, suppose the 90 percent of the population adheres to a fairness norm in which proposers make High offers and responders reject Low offers. A proposer who sticks to this norm will have an expected payoff of 2 euros. By switching to the unfair offer he would instead get 3 euros one time out of ten and nothing in the rest of the cases. By violating the norm her expected payoff drops to a meager 30 cents. By contrast, a responder who adheres to the norm and rejects unfair offers obtains 2 euros ninety percent of the times and nothing in the remaining 10% of the cases. By violating the norm and accepting all offers his payoff would increase of one euro with probability $\frac{1}{10}$, a mere ten cents in expectation.

This example contains an important message. When the fairness norm is followed by a large fraction of a population (for example when 90 per cent of the subjects make fair offers and reject unfair ones), proposers will be under a stronger pressure to learn that it pays to stick to it than responders are to learn that there is a profit to make in abandoning it. Intuitively, responders who respect the fairness norm are so slow to disappear that they drive to extinction proposers who make Low offers. When all offers are High, rejecting unfair offers has no cost, and the pressure against the norm disappears. The two populations may then remain locked for a long time in the set $\mathcal{N}$ in which the fairness norm is respected.

## 4 Social preferences

One of the earliest results in game theory is that backward induction implies that in the finitely repeated Prisoner's Dilemma the only subgame perfect Nash equilibrium is complete defection. In the 50.s mathematicians working at RAND corporation started running experiments to test whether this prediction was borne out by the data and obtained disappointing results. When the results were shown to John Nash he remained unimpressed. His reply was that game theory was not intended to be applied to such abstract situations.[6] In real life, games are rarely if ever played for a finite number of rounds known in advance by the players and it is unlikely that random people participating in an experiment could figure out the solution by themselves.

Nash's reaction is understandable, and in fact was very common among economists at the time, but it hides a trap. If any contrary evidence can be dismissed by saying that the subjects involved in an experiment did not understand the game, would not game theory become unfalsifiable? To avoid this conclusion, one must find at least a group of experimental results that would count as evidence against the predictions of game theory. A first hurdle in this direction is that game theory makes no prediction, unless one spells out the preferences of the subjects over the outcomes of a game. For example, an altruistic subject who maximizes a (weighted) sum of his and his opponent's monetary payoff is not irrational if he cooperates in the Prisoner's Dilemma. It would be more apt to say that he is not playing a Prisoner's Dilemma at all (Binmore, 1998a).

[6] I take this story from Holt (2007).

There are two routes one can take at this point. First, one may follow the traditional path of game theory in assuming that players are rational, but abandon the idea that they maximize their own monetary payoff in experimental settings. The literature explored several alternatives along this route, assuming that subjects could be inequity averse, reciprocally altruistic, guilt averse and so on.[7] Alternatively, one may retain the assumption that, by and large, individuals are selfish, and replace the idea of rationality with the more realistic alternative that they have to learn their way to an equilibrium. Note that, because of what we said in the previous section, the main difference between these alternatives lies not so much on the importance they put on other-regarding motives. The crucial difference concerns the type of equilibria one expects to emerge in experimental settings. In sticking to the rationality paradigm, the authors who developed the *social preferences* models endorsed all the notions of equilibria that had been developed in the refinements literature (see for example Battigalli and Dufwenberg (2009)). These notions had only to be adjusted to a context in which players were assumed to maximize more complex utility functions. By contrast, models based on learning were open to the possibility that a much larger set of Nash equilibria could be observed.

Being a firm believer in the revealed preferences approach, Binmore saw no problem in playing with different hypotheses about individuals' preferences to obtain better models of human cooperation. In fact, he repeatedly mocked scholars who invent the non-existent "self-interest axiom" that purportedly is at the basis of orthodox economics (Binmore, 2005). The social preferences approach is thus a perfectly legitimate scientific endeavour that sits firmly within the neoclassical tradition. However, this does not mean that it is also an interesting one. Whether the new models contributed in any way to our understanding of human cooperation revolved around an empirical question: Are there games in which the set of Nash equilibria for selfish players fail to predict subjects' behavior in controlled experimental situations?[8] Binmore's answer to this question was a resounding "no". There was no proof in the literature that, after an adequate time to familiarize with the game at hand, human players did not end up playing as if they had little or no concern for anything besides their own material payoff.

Binmore looked at the existing evidence using some well-respected summaries of the literature like Ledyard (1994), Sally (1995) and Camerer (2003). His conclusions can be summarized as follows:

- In the Public Goods Game (PGG) without punishment, cooperation is initially high but it declines with repetition and becomes close to zero.
- When a possibility of punishment is added to the PGG, cooperation remains high over time.

[7] The literature on social preferences is huge. An early survey of the literature is contained in Camerer (2003) while Cooper and Kagel (2011) contains a more up-to-date bibliography.

[8] Binmore spells out the condition that make a good experiment as (a) The game is simple, and presented to the subjects in a user-friendly manner, (b) The subjects are paid adequately for performing well, (c) Sufficient time is available for trial-and-error learning. See for example Binmore (2007).

- In the Ultimatum Game, fair offers do not decline over time and unfair offers are usually rejected.

Binmore concluded that this evidence cannot be used to mount a credible case against the Nash equilibrium hypothesis. The evidence shows that individuals are initially disposed to cooperate and behave fairly. However, cooperation and fairness declined over time in all the cases in which, because of lack of punishment, they did not correspond to an equilibrium. The evidence in the UG and in the PGG with punishment is not against Nash equilibrim. In both these games there are (non subgame perfect) equilibria in which players cooperate (in the PGG) and behave fairly (in the UG). So the existing evidence militate only against the notions of equilibrium that were invented in the refinement literature and that, Binmore insists, had been discredited by the evolutionary approach. Apparently, the social preferences literature had proposed a sophisticated solution to a non-existent problem.

There are at least two points that need to be stressed in this reconstruction of Binmore's position. First, learning models say nothing about subjects' motivations. In fact, when it comes to model the reasons behind subjects' choices, social preferences models are vastly more accurate than any model that couples bounded rationality with learning. If asked to motivate her choice, a subject who has just rejected an unfair offer in a UG will typically mention her own sense of fairness or her aversion to inequity. But from the shopkeeper's example in Sect. 2 we know that game theory is better used to explain the stability of a social norm, and not the explicit motives individuals have in following it.

Second, Binmore recognizes that learning models should not be used to make precise predictions about the outcome of any experiment.[9] In line with a long tradition of model building in economics, he pays little heed to the idea that a model can only be useful insofar as it can accurately represent an experimental phenomenon. As we saw in the previous section, learning models are useful because they attract scholars' attention on some aspects of an experiment that would otherwise pass unnoticed. In the study of the Ultimatum Game, for example, we discovered that in the vicinity of a non-subgame perfect Nash equilibrium one of the two players ceases to have any substantial incentive to make the right choice, because she is virtually indifferent among several alternative best replies. Her behavior would then be influenced by any aspect of the game that is being played and is not explicitly modelled.

The "rest of the profession" decided to follow another route.[10] For example, the leading explanation for the decline of cooperation in PGG was assumed to be the so called reciprocity hypothesis.[11] In two papers that generated a very large

[9] For example, "Nobody thinks that [the Replicator Dynamics] is anywhere near adequate by itself to predict how real individuals learn" Binmore (2010).

[10] "Experimenters responded to these [...] learning papers by largely ignoring them", (Cooper & Kagel, 2011). For a review of the literature see Chaudhuri (2011) and Fehr and Schurtenberger (2018).

[11] After an initial burst of enthusiasm for the explanations based on *pure* social preferences like inequity aversion or altruism, another strand of the literature emerged which stresses other motives behind

literature, Fischbacher et al. (2001) and Fischbacher and Gächter (2010) had elicited subjects' preferences for conditional giving in standard PGG. Subjects where asked how much they would contribute to the public good, if they were to choose *after* their group mates. These experiments revealed that over half of the subjects could be classified neither as selfish nor as altruistic. Rather, they were reciprocally cooperative: They were willing to contribute to the public good, as long as most of the others were doing their share. Pure, unconditional altruism was virtually non-existent while unconditional selfishness accounted for around 30 percent of the total. The reciprocity hypothesis' appeal stemmed from the fact that it could explain at the same time both the high initial level of contribution and its decline. Both phenomena were explained by subjects' evolving *beliefs* about the other subjects' behavior. The initially high contribution was explained by the fact that most subjects were best-responding to their belief that the other subjects would have contributed as well. The decline of cooperation that ensued was an obvious response to some of the other subjects' free-riding. But the real bonus of the reciprocity hypothesis was that it could explain a behavioral regularity that could not be explained in terms of learning, the so called "restart effect". Experiments had shown that if groups were rematched after a few rounds in which cooperation had declined almost to zero, in the new group most subjects started to cooperate anew, only to reduce their contribution again after a few rounds. It seemed as if players were constantly forgetting what they had just learned (Gintis, 2011).

A reader sympathetic with Binmore's positions may remain unconvinced. In a PGG that is repeated over time with a periodic reshuffling of the groups, the final payoff of a subject who contributes something in the early stages of each repetition differs for a few cents from the payoff of an unconditional defector because both subjects defect most of the times. As long as a group of subjects remains close to the non-cooperative Nash equilibrium for most of the rounds, the force with which learning pushes subjects towards the optimal decision is simply too small to remove all suboptimal choices.

A more impartial reader may get the impression that the debate has reached a standstill. Defenders of social preferences are right in pointing out that there are some phenomena, like the restart effect, that are hard to explain in terms of selfishness and learning alone. Binmore is right when he replies that social preferences can explain, at a cost of substantially more complex models, only small deviations from the predictions one could obtain by postulating pure selfishness. However, more recent experimental evidence has revealed that the role social preferences play in

Footnote 11 (Continued)

subjects' deviations from pure egoism. A prominent example in this direction is (Andreoni & Bernheim, 2009), in which pro-social behavior in experimental settings is explained in terms of a desire to preserve one's own social image in front of an audience. This type of pro-social behavior would disappear in a truly anonymous setting. Notice that this line of research is broadly in agreement with Binmore's position. An important difference, however, is that theoretical models of this type rely on the assumption that inexperienced subjects maximize a well-defined utility function, albeit more complex than the purely egoistic one which is common in standard economic models. Binmore instead believes that the attempt to rationalize the behavior of a subject playing a game for the first time is doomed to fail. (I thank an anonymous referee for attracting my attention on this issue.)

explaining cooperation in the PGG is probably even smaller than our reconstruction of the debate suggests. Burton-Chellew et al. (2016) elicited subjects' preferences for reciprocity using the procedure that is common in this literature. The novelty of their experiment is that they let part of the subjects play against computers that were programmed to generate random contributions. Surprisingly enough, even in this context a large fraction of the subjects revealed conditionally cooperative preferences. The authors conclude that conditional cooperators are confused about the situation they are in to the point that they are not able to tell the difference between playing against a human being or a computer.

The experimental results presented in Andreozzi et al. (2020) continue on the same vein. If subjects' deviations from self-interest is better explained in terms of a faulty understanding of the game, one should expect it to disappear after they gained some practice with the game at hand. To test this hypothesis, their experiment elicits subjects' preferences for conditional cooperation at every round of the game rather than just at the beginning, as it is common in the literature. Their data reveal that, although the majority of subjects can be classified as conditional cooperators in the early stages of the game, around 60 percent could be classified as purely selfish by the tenth round. They conclude that selfishness and learning explain a larger fraction of the decay of cooperation than supporters of the reciprocity hypothesis are willing to admit.

## 5 A Forerunner

In 2012 *Nature* published a short article titled "Spontaneous giving and calculated greed", (Rand et al. 2012), that was to have a large impact on the literature. The authors set themselves the task of providing a first characterization of the psychological determinants of human cooperation. The setting was the familiar dual-process framework in which decisions are the joint product of intuition and reflection (Kahneman, 2011). The authors claimed that "cooperation is intuitive because cooperative heuristics are developed in daily life where cooperation is typically advantageous". In the anonymous, one-shot social dilemmas subjects encounter in a typical experiment, however, these heuristics give the wrong advice because the cooperative action is not optimal. If subjects are given enough time to think about the game at hand, they will be more likely to suppress their cooperative impulse and choose the selfish, payoff maximizing option. According to the Social Heuristic Hypothesis (SHH),

> people internalize strategies that are typically advantageous and successful in their daily social interactions. They then bring these automatic, intuitive responses with them into atypical social situations, such as most laboratory experiments. More reflective, deliberative processes may then override these generalized automatic responses, causing subjects to shift their behavior towards the behavior that is most advantageous in context (Rand et al., 2014).

Binmore expressed very similar views in many different places. For example, in the review article he wrote for the influential collection of essays "Foundation of Human societies", (Henrich et al., 2004), we read:

> what should we expect to happen when we ask inexperienced subjects from small-scale societies to participate in a novel laboratory game designed to provide information on how people respond to situations involving social phenomena like fairness, trust, or reciprocity? The answer that seems obvious to me is that we should expect them to behave as they would behave in real life if they were offered similar cues to those offered in the laboratory. That is to say, we should use whichever equilibrium their own society operates in its repeated game of life to predict their initial behavior, rather than one of the equilibria of the one-shot game they are required to play in the laboratory. Binmore (2005)

He also stressed that the way a game is framed plays a crucial role, because different frames trigger different social norms.

> Habits are hard to shake off—especially if you are unconscious that you have a habit in the first place. So when the framing of an experiment triggers the appropriate environmental cues, we often respond with the habituated response: no matter how ill-adapted it may be to the actual game being played in the laboratory. Like a sailor stepping ashore, we still roll with the waves, even though there are no longer any waves with which to roll. I therefore think that Kahneman and Tversky's emphasis on the importance of framing in experiments is well grounded. (Binmore, 2007, 10)

Binmore never suggested that these (rather commonsensical) propositions could be used as a guide to make predictions about the behavior of inexperienced subjects playing experimental social dilemmas for the first time. Proponents of the SHH, instead, believed that this hypothesis had at least one empirically testable implication: since the intuitive choice in social dilemmas is to cooperate, one could induce subjects to choose more (less) selfishly by just giving them more (less) time to think. The initial experimental evidence the authors provided in support of this hypothesis was remarkably strong, which contributed to make it an instant success. In a recent survey of the literature, Capraro (2019) lists more than three hundred papers related to the SHH published in less than a decade.

Despite the air of familiarity between the SHH and the thesis that Binmore defended, there are also important differences. To begin with, the SHH emphasizes the role of the virtual form of learning that takes place within subjects' heads before the game is played (Weber, 2003), while Binmore favors the type of trial-and-error process that takes place in real time and requires repetition and experimentation.[12]

[12] This distinction is reminiscent of the old dichotomy between the *eductive* and the *evolutive* approach to equilibrium selection that Binmore introduced in Binmore (1987, 1988). In eductive models of equilibrium selection, e.g. in the *tracing procedure* introduced by Harsanyi and Selten (1988), the focus in on the way perfectly rational players would get to an equilibrium by simply thinking (assuming the pose of Rodin's *Thinker* (Binmore, 1998a, p. 87)) *before* a simultaneous moves game is played.

Another important difference stems from a more skeptical view that Binmore always had both towards the theoretical models social scientists use to capture reality and the experiments they run to falsify them. This aspect of Binmore's writings is more difficult to appreciate for readers who are less familiar with the philosophical parts of his production (Binmore, 1998a, 2005). While Binmore's contribution to formal economic modelling as published in regular scientific journals is barely distinguishable from similar products on the same topics, a more complex and nuanced picture emerges when these results are read on the backdrop of his works on the evolutionary foundation of the social contract. Although the first impression might suggest otherwise, a careful reading of these works reveal that Binmore has always been deeply aware of the limitations of current game-theoretical models. I shall illustrate this point with two examples. First, in his theory of the social contract Binmore placed a surprising little emphasis on the results obtained by the large literature on the repeated Prisoner's Dilemma that was started by Axelrod (2009).[13] The reason is that the fundamental result in the theory of repeated games, the so-called *folk-theorem*, proves that the intuition one can glean from the analysis of such a simple game is insufficient to do justice of the complexity of human cooperation. The folk-theorem shows that whenever a game is repeated over time a host of new equilibria emerge and players will in general have opposite preferences over some of them. Even the simple repeated PD has asymmetric equilibria in which (to borrow Binmore's favorite terminology) Adam cooperates once every two rounds, but expects Eve to cooperate at every round. This implies that every repeated game is at heart a bargaining game and that the equilibrium selection problem always involves an element of conflict among the players. Incidentally, this makes Binmore a precursor of the fairly recent literature on *extortion* in repeated games, that was started by the publication of Press and Dyson (2012). Their intriguing result shows that when playing a repeated PD a player may be tempted to obtain a larger payoff than his partner by playing what the authors call an *extortionate strategy*. The literature that it generated had started to explore the role of bargaining in repeated interactions, that is at the core of Binmore's social philosophy and had received relatively little attention in the literature. (Hilbe et al. (2015) contains a review of this literature.)

But even restricting the attention to the standard analysis of the repeated PD, Binmore's results show that there is no simple solution to the emergence of cooperation. His main result, contained in Binmore and Samuelson (1992), was that if agents have to pay a higher cost for using more complex strategies, then simple rules like Tit-for-Tat fail to pass even the Nash equilibrium test. The reason is straightforward: a strategy of unconditional cooperation is simpler than TfT but obtains, against TfT, the same payoff TfT obtains against itself. As a consequence, a population of players who use TfT would be invaded by unconditional cooperators that would in turn be invaded by defectors.[14] Binmore proved that evolution favors those strategies that *probe* the opponent in the initial stages of the game, to see whether it can be

[13] Axelrod's work was crucial in the generation of the very large literature on the repeated PD. However, Binmore's opinion was that Axelrod's work added little to what game theorist already knew at that time. See Binmore (1998b).

[14] Binmore worked on the tradition started by Abreu and Rubinstein (1988).

exploited by complete defection. It is only after a strategy proves that it can defend itself by responding with defection to defection that cooperation can start. This analysis casts doubts on the idea frequently repeated that evolution favors *nice* strategies that are never the first to defect.[15]

The considerations above should be sufficient to show that one should be extremely cautious when large generalizations about the purported intuitiveness of cooperation are made on the basis of simple evolutionary models. Maybe, as Binmore's results suggest, humans are hardwired to be cautious and distrustful when meeting a stranger, and the cooperativeness we observe in our societies even in early interactions with strangers is just the by-product of a stable institutional environment in which wrongdoers are routinely punished. Or maybe it is just the other-way round and human evolution has favored, as Axelrod (2009) suggested, *nice* strategies like Tit for Tat. As we saw in the previous section, Binmore never attempted to calibrate his rudimentary learning model on the available data and was ready to acknowledge that the few attempts that had been made in this direction gave disappointing results. Similarly, although he elaborated his own version of the SHH, he never presented it as a reliable guide to make predictions in experimental contexts. Just like it is hard to imagine a mathematical model that captures all the richness of real human learning in strategic settings, it is hard to imagine that the SHH can be a reliable guide to make predictions about the behavior of inexperienced subjects dealing with social dilemmas. Sometimes, one has simply to accept that there is a limit to the degree to which human behavior can be understood and predicted. The choices human subjects make when facing a game for the first time could be beyond this limit. This may help to explain why, after almost a decade of intense research, the verdict about the empirical relevance of the SHH is still pending (Camerer et al. 2018).

## 6 Conclusions

Binmore's many contributions to the literature on the emergence of human cooperation deliver two main messages, one negative the other positive. On the negative side, Binmore repeated relentlessly that it is pointless to adjust subjects' preferences to account for what is observed in laboratory experiments. The existing evidence is sufficient to show that although other regarding motives are certainly present, they are not strong enough to produces appreciable deviations from what would be observed if subjects were completely selfish. On a more positive note, Binmore maintained that the understanding of how cooperation is sustained requires the study of more complex games than the ones the literature has dealt with. When the game being repeated is asymmetric, for example, an equilibrium selection problem emerges on how the fruits of the mutually cooperative arrangements should be divided. Binmore was probably right in pointing out that social norms are more relevant as a mean to coordinate beliefs on different equilibria of a repeated game, rather than in creating new equilibria in one-shot games through other-regarding

[15] Holler and Klose-Ullmann (2020) contains a detailed analysis of Axelrod's experiment and a discussion of the lessons that Axelord believed could be learned from it.

motives like reciprocal altruism or inequity aversion. It is unfortunate that the elaborate solution he proposed for this problem in his more philosophical works attracted so little attention. It may offer some consolation the fact that his ideas occasionally resurface in the literature, although Binmore's name hardly gets a mention. It would not be surprising if, in the due time, scholars will rediscover one by-one-the many gems that are scattered in the large *corpus* of his works.

**Acknowledgements** I thank Ken Binmore, Manfred J. Holler and two anonymous referees for their comments on a previous version of this note.

## References

Abreu, D., & Rubinstein, A. (1988). The structure of Nash equilibrium in repeated games with finite automata. *Econometrica, 56*(6), 1259–1281.

Andreoni, J., & Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, *77*(5), 1607–1636.

Andreozzi, L., Ploner, M., & Saral, A. S. (2020). The stability of conditional cooperation: Beliefs alone cannot explain the decline of cooperation in social dilemmas. *Scientific Reports, 10*(1), 13610.

Axelrod, R. (2009). *The evolution of cooperation: Revised edition*. New York: Basic Books.

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory, 144*(1), 1–35.

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

Binmore, K. (1998a). *Game theory and the social contract, Volume 2: Just playing*. The MIT Press.

Binmore, K. (1998b). Review of R. Axelrod, the complexity of cooperation: Agent-based models of competition and collaboration. *Journal of Artificial Societies and Social Simulation*, **1**(01).

Binmore, K. (1987). Modeling rational players: Part I. *Economics and Philosophy, 3*(2), 179–214.

Binmore, K. (1988). Modeling rational players: Part II. *Economics and Philosophy, 4*(1), 9–55.

Binmore, K. G. (1990). *Essays on the foundations of game theory*. B. Blackwell.

Binmore, K. (2005). Economic man or straw man? *Behavioral and Brain Sciences, 28*(6), 817818.

Binmore, K. G. (2005). *Natural justice*. Oxford University Press.

Binmore, K. (2007). *Does game theory work? The bargaining challenge* (1st ed., Vol. 1). The MIT Press.

Binmore, K. (2010). Social norms or social preferences? *Mind & Society: Cognitive Studies in Economics and Social Sciences*, *9*(2), 139–157.

Binmore, K. (2020). *Crooked thinking or straight talk?: Modernizing epicurean scientific philosophy*. Springer International Publishing.

Binmore, K. G., & Samuelson, L. (1992). Evolutionary stability in repeated games played by finite automata. *Journal of Economic Theory, 57*(2), 278–305.

Binmore, K., & Samuelson, L. (1994). Drift. *European Economic Review, 38*(3–4), 859–867.

Binmore, K., & Samuelson, L. (1994). An economist's perspective on the evolution of norms. *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft, 150*(1), 45–63.

Binmore, K., & Samuelson, L. (1999). Evolutionary drift and equilibrium selection. *The Review of Economic Studies, 66*(2), 363–393.

Binmore, K., Swierzbinski, J., & Proulx, C. (2001). Does minimax work? An experimental study. *The Economic Journal, 111*(473), 445–464.

Bowles, S., & Gintis, H. (2013). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.

Burton-Chellew, M. N., Mouden, E., Claire, & West, S.A. (2016). Conditional cooperation and confusion in public-goods experiments. *Proceedings of the National Academy of Sciences, 113*(5), 1291–1296.

Camerer, F.C., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, A.B. Nosek, A.B., Pfeiffer, T., Altmejd, A., Buttrick, N., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., & Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press Princeton.

Capraro, V. (2019) The Dual-Process Approach to Human Sociality: A Review. Available at SSRN: https://ssrn.com/abstract=3409146 or https://doi.org/10.2139/ssrn.3409146

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics, 14*(1), 47–83.

Cooper, D. J., & Kagel, J. (2011). Other regarding preferences: A survey of experimental results. In J. Kagel & A. Roth (Eds.), *The handbook of experimental economics* (Vol. 2). Princeton University Press.

Cressman, R., & Schlag, K. H. (1998). The dynamic (In)stability of backwards induction. *Journal of Economic Theory, 83*(2), 260–285.

Elster, J. (1979). *Ulysses and the Sirens: Studies in rationality and irrationality*. Cambridge University Press.

Elster, J. (1989). *The cement of society: A survey of social order*. Studies in rationality and social change. Cambridge University Press.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour, 2*(7), 458–468.

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *American Economic Review, 100*(1), 541–56.

Fischbacher, U., Gachter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters, 71*(3), 397–404.

Friedman, D. (1991). Evolutionary games in economics. *Econometrica, 59*(3), 637–666.

Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games*. EBSCO eBook Collection. MIT Press.

Gale, J., Binmore, K. G., & Samuelson, L. (1995). Learning to be imperfect: The ultimatum game. *Games and Economic Behavior, 8*(1), 56–90.

Gintis, H. (2011). Reply to Binmore: Social norms or social preferences? Available at http://www.umass.edu/preferen/gintis/ReplyToBinmore.pdf

Gintis, H., Bowles, S., & Boyd, R. T. (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. The MIT Press.

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

Henrich, J. P., Boyd, R., Bowles, S., & Fehr, E. (2004). *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford University Press.

Hilbe, C., Traulsen, A., & Sigmund, K. (2015). Partners or rivals? Strategies for the iterated prisoner's dilemma. *Games and Economic Behavior, 92,* 41–52.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press.

Holler, M. J., & Klose-Ullmann, B. (2020). *Scissors and rock: Game theory for those who manage*. Springer International Publishing.

Holt, C. A. (2007). *Markets, games, and strategic behavior*. Pearson Addison Wesley. Addison-Wesley series in economics.

Kahneman, D. (2011). *Thinking*. Fast and slow. Farrar, Straus and Giroux.

Ledyard, J. (1995). Public goods: A survey of experimental research. In J. Kagel & A. Roth (Eds.), *Handbook of Experimental Economics*. Princeton University Press.

Press, W. H., & Dyson, F. J. (2012). Iterated Prisoner's Dilemma contains strategies that dominate any evolutionary opponent. *Proceedings of the National Academy of Sciences, 109*(26), 10409–10413.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*(Sep), EP 427.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5*(1), 3677.

Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society, 7*(1), 58–92.

Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. Economic learning and social evolution, The MIT Press.

Smith, J. M. (1982). *Evolution and the theory of games*. Cambridge University Press.

Weber, R. A. (2003). 'Learning' with no feedback in a competitive guessing game. *Games and Economic Behavior, 44*(1), 134–144.

Weibull, J. W. (1997). *Evolutionary game theory*. The MIT Press.