Modeling *homo sociologicus*: social influence and interdependent behavior in economics

Michiru Nagatsu

January 26, 2021

1 Introduction

Economics has been criticized by other disciplines, and by some prominent economists themselves, for many of its methodological choices, in particular the specific assumptions in models. Some of these critiques have been appropriated by economists, giving rise to new sub-fields of economics. Critics from the field of history and sociology have long pointed out that the economic models of markets are not historically and institutionally embodied, as the real markets are. Various strands of institutional economics arose from such criticism. Claims by ecologists that natural capital is not completely substitutable with other forms of capital fostered the development of ecological economics. Likewise, criticism from psychologists that decision makers are irrational or boundedly rational foster the development of behavioral economics. This chapter concerns another example of persistent criticism, namely the assumption in economics that *homo economicus* is an asocial animal, which does not hold on evolutionary, psychological or sociological grounds. How have economists responded to this criticism, and how successful have they been?

To focus the discussion, I will review four strategies of model modification that aim to address the sociality of human behavior and interactions in the framework of rational choice. In particular, I will focus on modifications of game theory, which is the main analytical tool economists use to model interdependent choices. The perspective I adopt here reflects mainstream approaches to economic modeling, and I do not discuss more radical proposals to use entirely new frameworks such as complexity theory or network theory (see e.g. Mason et al., 2007). This focus on the mainstream approaches is motivated as a useful way to highlight core characteristics of economic explanations of sociality. The review of four major approaches to modeling social influence in economics will indicate different interpretations of sociality, followed by a suggestion that the non-reductive and constructive interpretation brings economics closer to sociology and away from psychology and biology.

The chapter is organized as follows. In the next section, in order to make explicit the explanatory core of economic models, I present a broad categorization of explanations of human cooperation, building on Calcott (2013), and specify the category in which economic explanations fall. After that, I describe in some detail four types of responses to the question of asociality. In section 4, I discuss two methodological issues. The first concerns disagreements about what these models are models *of*, which reflects the core explanatory interests of economics discussed in Section 2, and the second type concerns the practical use of alternative models of socially influenced behavior. Section 5 concludes the chapter.

2 Explanations of social behavior: carving the domain of economics

To discuss the domain of economics, at least two approaches are possible. One is to start from the debates about the proper explanandum of economics, e.g. whether it should explain market phenomena narrowly following Coase or social phenomena more broadly following Becker (Hsiung, 2001). The other approach, which I adopt here, is to focus on explanans, namely types of explanations and identify core features of economic explanations in contrast to other types. Let us focus on the observation of pro-social human behavior broadly construed, which is an uncontroversially important explanandum of economics. There are many factors contributing to human cooperation on different levels from the neural and the cognitive to the affective and the social. The standard explanatory division of labor distinguishes the proximate from the ultimate causes of human sociality in pro-social or cooperative behavior (Mayr, 1961). The proximate cause is associated with the psychological mechanism of social behavior, and the ultimate cause with the evolutionary dynamics that give rise to such behavior. In this sense, the division appears to be between psychology and biology, and there is no place for economics. However, I will show that the proximate-ultimate dichotomy is a good starting point to understand distinct features of economic explanations, drawing on Calcott's (2013) modification of the dichotomy.

In the context of non-human cooperation in biology, Calcott (2013) criticizes this dichotomy for obscuring two distinct dimensions along which different types of explanations can be distinguished.[1] The first dimension is temporal. Proximate explanations refer to mechanisms, typically psychological, that support currently observed cooperative behavior such as a sense of duty and sympathy. Referring to these mechanisms, however real they may be, does not explain how it was possible for

[1]In what follows I only discuss human cooperation, although the scope of Calcott's discussion is wider.

them to evolve, given the background knowledge that cooperative behavior does not always enhance individual fitness. To address this explanatory demand, one needs to turn to evolutionary explanations that explicitly model population dynamics over an evolutionarily relevant period of time, such as gene-culture co-evolution, group-selection, and so on. In other words, this type of explanation addresses the historical or evolutionary origins of the explananda.

These two types of explanation also differ along another dimension, the spatial scale of resolution. Whereas proximate explanations focus on mechanisms operating on the individual (organism) level, ultimate explanations model the interactions of individuals an the population (organization) level (see Table 1).

Table 1: Two dimensions of the proximate-ultimate distinction. Reproduced from Calcott (2013, 253)

	Current Operation	Historical Origins
Individual Mechanisms	Proximate explanation	*Lineage explanation*
Population Dynamics	?	Ultimate explanation

Calcott's main purpose in disentangling these temporal and spatial dimensions of the proximate-ultimate distinction is to highlight the distinct third category, which he calls *lineage explanations*. Such explanations focus on historically long periods of time, such as the evolution of the vertebrate eye or the shape of a feather. Keeping natural selection on the population level as a background assumption, they show how step-by-step variation can produce a complex camera eye from a simple eye spot, for example. The focus is on the individual, but the time span is evolutionary, a continuous trajectory of change over time. The explanatory interest lies in how new, complex capacities or forms were able to emerge from pre-existing and relatively simple forms. Calcott goes on to argue that the complex cooperative patterns of organizations also call for lineage explanations. He provides a few examples of biological cooperation for which lineage explanations are needed, such as the cooperative nest-building of green tree ants, team hunting by mammals, and the internal organization of differentiated cells into a useful structure. All of these examples involve a complex division of labour and the organization of different roles into a new function.

Analogous explanations abound in history and the social sciences. Although the focus is on social and institutional rather than biological organizations, and the temporal scale is human historical rather than evolutionary, Max Weber's idea that the roots of modern capitalism lie in Protestant ethics, or rather in law (Tigar and Levy, 1977), fall into this type of the new-from-old explanation. The explanation of societal transitions lie in the re-deployment of existing institutions, values, norms and so on for new purposes.

What is peculiar about Calcott's table, however, is that the box corresponding to current operation/population-dynamics is left empty. This is presumably because population dynamics in biology refer to the dynamics of natural selection over an evolutionary temporal scale. However, just as lineage explanations correspond to specific explanatory interests in mechanistic step-by-step changes through variations in some feature of an organism on the individual level, there is an explanatory interest in the current operation of some feature on the aggregate level. In contrast to population dynamics, the focus is on the current operation of some causal processes that give rise to a phenomenon that is being observed; and in contrast to individual mechanisms, the focus is on patterns of interaction among individuals.

My suggestion is that economic explanations typically fall into this category. For example, economic explanations of cooperation (or the lack thereof) keep individual mental operations as background assumptions and model current operations as the aggregate result of the interactions of individuals.

In particular, equilibrium explanations that are typical of microeconomics satisfy the conditions of being population-level and a-temporal (non-dynamic). If this is the typical mode of economic explanations, many of the idealizations economists deploy in their modeling make a lot of methodological sense. The assumption that individual actors have narrow self-interests, for example, is merely a background condition that fixes one parameter to see how their interactions give rise to population-level phenomena. Similarly, the assumption that all individuals are the same, whether selfish or not, is an idealization that helps focus on population-level interactions. In fact, as I will show in the next section, most well-known models of social preferences could be interpreted as adjustments of background assumptions, keeping the core explanatory device of equilibrium analysis unchanged. Moreover, this four-fold scheme enables us to highlight the novelty of more recent models of social behavior, while at the same time showing them as distinctly economic.

3 Models of socially influenced behavior

Let us come back to the criticism that *homo economicus* is asocial, and the responses to it. Some influential models of pro-social behavior appeared already in the 1980s, including those of Gary Becker and Robert Sugden. This suggests that on the methodological level it was relatively straightforward to include other-regarding preferences as an independent variable of the decision maker's utility function. However, empirical evidence of pro-social behavior for these models was in the form of aggregate observational data, making their empirical evaluation somewhat complicated.

The literature on social preferences exploded after the invention of a series of two-person games, namely the ultimatum game, the dictator game and the trust

game (Camerer, 2003), which encouraged economists to propose a variety of models of social preferences (Fehr and Schmidt, 1999; Rabin, 1993) and their systematic empirical evaluation.

One way of categorizing different models of social behavior is to distinguish between models of preferences and models of beliefs or reasoning. I will follow this strategy, although some models do not fit in this dichotomy, nor it is clear that economists are committed to belief-desire psychology, as we will see below.

Two separate assumptions in models of preferences are targeted by critics. The first is that people are selfish, or in more neutral language, self-regarding, meaning that they care only about their own payoffs. The second is that people's preferences are independent of each other. These are, in fact, distinct points, which have been addressed by distinct models. The other set of criticisms concerns strategic rationality, in particular best-reply reasoning as an inaccurate description of how people reason in interdependent situations (Section 3.2).

3.1 Preference-focused approaches

Given the methodological analysis of explanatory styles and corresponding modeling strategies across disciplines, it is unsurprising that the first response of economists was to modify the selfishness assumption, as shown in Equation 1.[2]

$$U_i = f(x_i) \tag{1}$$

where U_i is the utility function of individual i, and x_i is the payoff,[3] and f is some monotonically increasing function. It is unsurprising because such an assumption is simply a background assumption that is not essential to the core explanatory project. There are good overviews of this literature available (e.g. Dhami, 2016; Camerer, 2003), so I will keep my presentation here brief and schematic. In particular, I will not engage in systematic evaluations of the empirical success of different models against the data from a range of experimental games. My aim is rather to highlight the differences in approach to issues of socially influenced behavior from a methodological perspective.

The first knee-jerk reaction of economists to the criticism that people are not selfish is to propose *altruism* as an alternative motivation. This approach is intuitive because the contrast target is egoism or selfishness. It is also natural, because the salient phenomenon to be explained is the field observation that many public goods are financed through private contributions to charities such as the Red Cross. People

[2]It even precedes the explosion of the social-preference models after the invention of the utlimatum (Güth et al., 1982) and dictator (Kahneman et al., 1986) games. See e.g. Becker (1974).

[3]In what follows, I use the term payoff, although terms such as consumption and income are used variably in economics, depending on the context.

privately contribute to the public goods through charitable giving, which is a puzzle that the simple assumption of selfishness cannot explain.[4] Two types of altruism have been distinguished: one is pure and the other is impure, also known as 'warm glow' altruism (Andreoni, 1990). Both are described as the following utility function:

$$U_i = f(x_i, G, g_i), i = 1, ..., n. \tag{2}$$

where

$$G = \sum_{i=1}^{n} g_i$$

G is the total amount of public good and g_i is i's contribution. A model of pure altruism does not contain g_i in the utility function: the agent cares only about private income (x_i) and the total amount of public good. In contrast, impure altruists also care about how much they contributes to the public good as such, independently of how that contributes to the total public good. In the extreme case G is removed from the model, making the altruist completely egoistic (in the sense of caring only about his or her own altruistic act regardless of its impact on the public good).

The model of impure altruism is introduced to account for several stylized observations that models of pure altruism cannot explain, such as the fact that government grants do not crowd out private giving.[5] Although this model has been fairly popular in economics, it is not considered 'modern', partly because it is some kind of regress to egoism in introducing *more* self-regarding preference to account for prosocial behavior.

Many of what Camerer (2003, 101) calls modern theories, in contrast, 'substitute a social utility for a vector of payoffs.' In other words, what people care about is their relative as well as absolute payoffs. This methodological choice allows economists to model the psychological mechanism of *interpersonal comparison* based on some notion of fairness, as a main channel through which social influences modify individual behavior. Significantly, a range of two-person experimental games enabled the operationalization of interpersonal comparison, as opposed to focusing on the aggregate of G or one's contribution g_i. The most well known of these is a model of inequity aversion (Fehr and Schmidt, 1999), which includes the difference (both positive and negative) between i's and the reference person j's income as sources of

[4]Because most experimental subjects in a neutrally framed public-good game turn out to be either unconditionally selfish or conditionally cooperative, one might think that economists have poor psychological intuition. However, it is worth pointing out that the explanandum was charitable giving in the field, not a voluntary contribution to the public goods in the lab.

[5]Crowding out is predicted because when G becomes bigger by means of government grants, i should now allocate more money to her own private good x_i to maximize her overall utility. See Becker (1974, Section 3.B, in particular footnote 34).

disutility.[6]

$$U_i = x_i - \alpha \max[x_j - x_i, 0] - \beta \max[x_i - x_j, 0] \tag{3}$$

In words, i's utility increases as her income does, but decreases when the reference person (j) earns more than she does, factored by α, or less than she does, factored by β. It is usually assumed that $\alpha > \beta$, meaning that i's hatred of the inequity whereby she earns less than j is stronger (due to envy) than her hatred of the inequity whereby she earns more than j (due to guilt or pure altruism).

Another well-known model is that of reciprocal fairness (Rabin, 1993). The psychological intuition behind it is that people like to reciprocate what they perceive as kind behavior with a kind response, and what they perceive as unkind or mean behavior with an unkind response. It thereby captures a belief-dependent preference, which is distinct from the preference for equality of outcomes *per se* captured by models such as that of Fehr et al. (1993). Rabin (1993) uses a technique called *psychological game theory* to model this. The utility function is represented as

$$U_i(a_i, b_j, c_i) = \pi_i(a_i, b_j) + \alpha \tilde{f}_j(b_j, c_i) + \alpha \tilde{f}_j(b_j, c_i) \cdot f_i(a_i, b_j)$$

(4)

The first term $\pi_i(a_i, b_j)$ captures i's monetary payoffs determined by her own action (a_i) given her belief about j's action (b_j); the second term $\alpha \tilde{f}_j(b_j, c_i)$ captures how much i care about j's perceived kindness toward her, based on b_j (her belief about j's action) and c_i (her belief about j's belief about her action); and the third term $\alpha \tilde{f}_j(b_j, c_i) \cdot f_i(a_i, b_j)$ captures i's reciprocal preference as a product of her perception of j's kindness (the same as the second term) and her own kindness ($f_i(a_i, b_j)$), determined by a_i (her own action) and b_j (her belief about j's action). Finally, α represents how much i cares about j's kindness and the importance of reciprocation, relative to monetary payoffs. It is straightforward that you feel good when others are nice to you (the second term), but the crucial mathematical trick in this model is the third term, which is positive whenever you reciprocate (kind to kind or mean to mean). i's own kindness is determined by how much j receives relative to a pre-determined fair point on the scale of maximum and minimum payoffs. For example, if 50-50 is the fair point in an ultimatum game, i's kindness is positive whenever she offers more than that to j.

This model of reciprocal fairness defines a new equilibrium concept, *fairness equilibrium*, in which players maximize the utilities defined above, and their beliefs

[6]The following equation is a two-person model to communicate the main idea, but the original is a general n-person model.

are correct about the behavior and beliefs of others, namely, $a_i = b_j = c_i$ (similarly $a_j = b_i = c_j$). Camerer (2003, 113, 116) summarizes a comparison of the three types of models described above:

> Altruism theories do not explain both negative and positive behavior toward others without crudely changing the signs of coefficients exogeneously. [Comparing the inequality-aversion models and the reciprocal-fairness models,] The reciprocity-based view is surely more psychologically correct because players do care about the intentions of other players and unchosen paths. At the same time, inequality-aversion is easy to use analytically because social utilities can just be substituted into cells of a payoff matrix, or nodes of a tree, before doing standard equilibrium analyses.

In this quote Camerer (2003) mentions two factors, in addition to empirical success, that are relevant for the methodological evaluation of these other-regarding preference models. One is psychological realism, and the other is ease in modeling. But Camerer also suggests that models may not converge in the long run, noting that 'when the empirical dust settles, both approaches may prove useful in different technical applications' (Camerer, 2003, 113). Interestingly, this quote suggests a sense in which economics is more like applied engineering science, rather than basic behavioral science, which we will discuss later. Before going there, however, let us discuss in the next subsection what Camerer misses in his assessment, namely the problem of how to model norm-based behavior.

3.2 Norm-focused approaches

The problem of both models of social preferences is that the fairness reference points that are crucial for inferring inequality or others' intentions are exogenously given (see Dhami, 2016, 439-440). In reality, however, what is fair depends on the context in the sense that it gives players an appropriate notion of fairness at first, and they play accordingly. Empirically, this is an under-determination problem: when the economist observes some data and tries to estimate a model, should she adjust parameters such as α and β, or the value of the fairness reference point? Where does the latter come from? It is less problematic when the observer or experimenter is confident that she is controlling for the reference point because she is an insider to that population, but the problem is severe when she has to explain cross-cultural variations in terms of what norms apply in what situation (Gurven, 2004, 225).

On a more abstract level, Dhami (2016, 963) summarizes the problem in the following passage:

> Classical game theory is based on *methodological individualism*. By contrast, the epistemic conditions for a Nash equilibrium require common priors and common knowledge, which appear to transcend personal characteristics of individuals. Indeed, these constructs appear to depend on social institutions that help align the beliefs and expectations of individuals.

Note that refining Nash equilibrium as a fairness equilibrium as Rabin (1993) does fails to address this problem: if anything, it exacerbates it in making the epistemic conditions even more unrealistic.

With a view to reconciling this problem with a formalism of game theory, Gintis (2009) proposes using *correlated equilibrium* in epistemic game theory[7] explicitly to incorporate the roles that social norms and conventions play in coordinating the expectations and behavior of individuals. A figurative person called *choreographer* moves first in an augmented game, giving recommendations as to what strategies players should adopt. A correlated equilibrium is a Nash equilibrium in that no player can do any better than follow the choreographer's recommendation, given that the others do the same. Gintis assigns the role of choreographer to conventions and social norms.

This conceptual move assumes the main task of social norms to be that of coordinator in games with multiple equilibria, akin to the function of focal points or salience (Schelling, 1960). However, social norms should play a more extensive role than facilitating coordination if they are to explain empirical anomalies of non-equilibrium play in, for example, mutual cooperation in finitely repeated social-dilemma or public-good games, or other observations from two-person sequential games.

In fact, Gintis (2017, chapters 6 and 10) proposes that a minority of human populations have internalized 'strong reciprocity,' an altruistic trait of following norms unconditionally and punishing violators at a personal cost. Although the majority are self-regarding, the sufficient chance of encountering strong reciprocators deters them from norm deviation, thus sustaining norm-following as an aggregate stable phenomenon.

An alternative approach to modeling the extra-coordinative roles of social norms in influencing aggregate behavior is to see them as self-sustaining or reinforcing mechanisms of *conditional* preferences, as Bicchieri (2017, 35) notes:

> A social norm is a rule of behavior such that individuals prefer to conform to it on condition that they believe that (a) most people in their reference network conform to it (*empirical expectation*), and (b) that most people

[7]Epistemic game theory models the structure of knowledge in normal form games by introducing information, subjective priors and conjectures of players about which pure strategies others play.

in their reference network believe they ought to conform to it (*normative expectation*).

Whereas for Gintis the key mechanism is the minority's unconditional (internalized) preference for compliance, for Bicchieri it is the majority's preference for compliance conditional on the majority's compliance. Basic preferences for conforming to others' empirical and normative expectations are presupposed (with individual differences), which is compatible with their having some evolutionary roots and having been internalized through socialization, as Gintis theorizes about strong reciprocity. However, such a preference is assumed to be common although weak, in the sense that it influences behavior only in the event of widespread conformity and normative expectations. Another difference is that Bicchieri's model of conditional preferences aims to explain and facilitate rapid societal behavioral changes, whereas Gintis seems to be more interested in explaining how strong reciprocity plausibly evolved over a much longer time-scale.[8] In other words, Gintis's model is located more toward the bottom right-hand section in Table 1.

There are lively discussions about the empirical success of these models (see e.g. Guala, 2012, and the related commentaries), but what is important to note here is the common feature of both accounts, namely that they do not simply re-define the objective functions of representative players, and solve the new game by means of Nash or a modified equilibrium.[9] Instead, both accounts explicitly hypothesize how people exert influences on each others' behavior. Gintis models the stability of norm compliance as a consequence of having a mix of people with heterogeneous preferences,[10] whereas Bicchieri relies on belief-desire psychology and the related networks (i's behavior sends information about her beliefs and preferences to j, who updates his beliefs and behavior, and so on) to model dynamic processes of influence propagation in the aggregate. The level of conformism in this model is allowed to vary within the population (i.e., different people may have different levels of conformist preferences), but it is assumed to be stable for any given player. On the other hand, influence propagation is mediated by belief rather than preference changes. For example, compliance among hardline conformists could be perceived as evidence of compliance by moderate conformists, who then conform, which could,

[8]For example, Gintis (2017, 242) discusses the population dynamics through which strong reciprocity could have evolved by comparing a genetic-group selection model and a gene-culture co-evolutionary model.

[9]Bicchieri (2006, Appendix of chapter 1) does this exercise in much the same way as the other modelers of social preferences, but she does not seem to take it as central in her account.

[10]This is distinct from evolutionary game theory, in which strategies as behavioral phenotypes, not human agents, play games and occupy evolutionary niches, for example tit-for-tat in the repeated prisoner's dilemma dominates the population through replicator dynamics, or different strategies share portions in evolutionary equilibrium. Gintis's claim is instead that norm-following dominates behaviorally, but is backed-up by populations with different preferences and beliefs.

in turn, be perceived as evidence of compliance by weak conformists, and so on until even the weakest conformists are convinced that there is reason to comply. Norm-following behavior could also unravel through this chain of influence propagation in a similar way. The model also explains the phenomenon of *pluralistic ignorance*, according to which an unpopular norm (such as female genital cutting or college binge drinking) that is unsupported according to people's private preferences may continue to be followed because people mistakenly perceive any observation of the majority's compliance as evidence of collective support.

3.3 Models of conditional preferences

Whereas Gintis assumes that norm-following preferences are intra-personally stable as a result of evolutionary and cultural processes, recent models purport explicitly to model the phenomenon according to which preferences are directly influenced or conditioned by others' preferences within a relatively short time scale. One could interpret this as a general approach to conditional preferences that is compatible with Bicchieri's belief-desire model of preferences that are conditional on expectations.

Commenting on the type of social-preference models reviewed in Section 3.1, Stirling and Felin (2013, 2) note that 'the payoffs associated with these approaches are explicitly categorical, and any sociality generated by these models remains a function of individual interests.' The underlying criticism in this passage is sociological, similar to Dhami's (2016) claim that individual beliefs are a function of socialization. In the words of Fershtman and Segal (2018, 127), the model has explicitly to focus on 'the formation of endogenous behavioral preferences that are subject to social influence.' The term *social* signifies two things. First, this approach is distinct from evolutionary game theory, in which relative success or fitness is exogenously given according to some survival standards, economic (monetary pay-offs) or evolutionary (reproductive success). Second, one's preferences are influenced not only by parents, as in cultural transmission models, but also by other members of the social group. In other words, this approach is located in the bottom-left section of Table 1, namely within the core explanatory interests of economics.

Fershtman and Segal (2018) propose that individuals have two distinct utility functions, u_i representing their core preferences, and v_i their behavioral preferences, which are related as follows:

$$v_i = f_i(u_i, v_{-i}), \tag{5}$$

where v_{-i} represents the behavioral utility profile of $n-1$ individuals in the social group (everyone except i). In other words, i's behavioral utility is influenced by the behavioral utilities of other members and her own core utility, via some *social influence function* f_i. In particular, they impose a simplification on this function,

namely that v_i is a function of her core preferences and the *average* observed behavioral preferences of everyone else. This assumption simplifies the influence functions as if there were only two persons, but at the same time it allows preference *distribution* to have an impact on equilibrium behavioral preferences (behavioral preferences after they have been influenced by each other). That is, individuals may form different behavioral preferences by observing different subsets of all behavioral preferences. On the basis of this and other assumptions, Fershtman and Segal (2018) prove several things, such as the existence of a unique equilibrium in which behavioral preferences are stable; that the influence propagation does not reverse the core preference patterns (e.g. if i is more risk-averse than j at the core, then this order is preserved in behavioral preferences after the influences have been exerted); and that the influence does not make individuals more extreme (e.g. polarization does not happen).

Given that the proposed model is still at the stage of conceptual formalization and there are no empirical applications, it is difficult to evaluate it thoroughly. For present purposes, however, it is worth noting that Fershtman and Segal (2018) explicitly purport to model social influences upfront, while remaining neutral with respect to psychological mechanism. Although they consider psychological and physical constraints on how sensitive individuals can be to changes in outcomes, they explicitly refrain from supporting varying psychological concepts such as altruism, reciprocity, and group identity. Moreover, they remain open as to *the reason why* other people's behavior affects one's own behavior. In fact, even basic conformist preference, which is a necessary assumption for Bicchieri and Gintis, is not assumed. It is revealing in this respect that they claim their model can be applied to the analysis of committee deliberation: 'The effect of deliberation can be captured by our social influence procedure where each individual votes according to his behavioral preferences which depend on his core preferences and the behavioral preferences of committee members that participate in the deliberation. [...] Adopting our setup, the different procedures may affect the formation of the behavioral preferences and therefore the outcome of the committee's vote.' (Fershtman and Segal, 2018, 140)

Similar interests are salient in the model of *conditional game theory* (Stirling, 2012; Stirling and Felin, 2013; Hofmeyr and Ross, 2019; Ross and Stirling, 2020). For example, the main concern of these theorists is not to move the study of preference formulation to 'the psychological and sociological headwaters of preference origination' (Stirling and Felin, 2013, 2), but rather to provide a general framework that is applicable to 'a wide range of potential social contexts that feature extant social relations and influence.' Their motivating examples, such as strategic decisions by a Board of Directors (Hofmeyr and Ross, 2019) and manager-employee relations (Stirling and Felin, 2013), are thus similar to Fershtman and Segal's (2018) committee deliberation. Some context-specific social relations are already assumed in all these

examples, but they are not tied to specific preferences coming from background assumptions given by evolutionary or psychological theories. In other words, they are compatible with heterogeneous explanations of social influences, such as that individuals 'may like (or dislike, for that matter) the others involved, they may value others' opinions, or they may have an existing relationship with others (familial, friendship or professional).' (Stirling and Felin, 2013, 2)

The modeling approach of Stirling and others is rather different from that of Fershtman and Segal (2018), however. Rather than assuming core and behavioral preferences for each players, the assumption in conditional game theory is that preferences may be *categorical* or *conditional*. The former are normal preferences as assumed in classical game theory, whereas the latter are conditional on other players' preferences. Suppose, for example, that there are two players, e (employee) and m (manager), with two feasible actions for each, h_e, l_e and h_m, l_m, corresponding respectively to high and low employee work effort and the monitoring efforts of the manager. The players have preferences over the outcome space $\mathcal{A}_e \times \mathcal{A}_m$, which contains four action profiles $(h_e, h_m), (h_e, l_m), (l_e, h_m), (l_e, l_m)$. In classical game theory, both e and m have categorical preference orderings, such as $(l_e, l_m) \succ_e (h_e, h_m) \succ_e (h_e, l_m) \succ_e (l_e, h_m)$ for e, and $(h_e, l_m) \succ_m (h_e, h_m) \succ_m (l_e, h_m) \succ_m (l_e, l_m)$ for m. In contrast, the influence m has on e is modeled in conditional game theory as e's conditional preference orderings: e entertains hypothetical propositions in the form of '*if* m's preference ordering is such and such, *then* e's preference ordering is such and such', and so on. The antecedents of these hypothetical propositions are called *conjectures*, and their consequents are *conditional utilities*. A conditional utility of e is defined as

$$u_{e|m}(\cdot|\boldsymbol{a}_e) : \mathcal{A}_e \times \mathcal{A}_m \to \mathbb{R} \tag{6}$$

for each conjecture $\boldsymbol{a}_e$. Such utilities are non-negative and sum to unity, but no more specific constraints are imposed on them based on psychological hypotheses. On the contrary, this mathematical modeling framework is meant to be as general as possible, while providing a way to model conditional preferences analogous to conditional probabilities or beliefs in multivariate probability theory. Conditional utilities have all the properties of probability mass functions. In addition, with technical conditions (acyclicity and framing invariance[11]), conditional preference

[11]Framing invariance means that informationally equivalent but notationally different framings of preference aggregation will result in the same joint preference ordering. Stirling and Felin (2013, 4) note that 'The richness and variability of human behavior, however, make it impossible to impose this condition without justification', and justify it as a weaker condition than the assumption of classical game theory that preferences are categorical (and therefore framing invariant). Acyclicity means that influence flows must be unidirectional, e.g. it cannot be that m influences e, which in turn influences m, leading to infinite regress. Ross and Stirling (2020) note that 'In many social settings, however, social relations are cyclic', and extend the frame to accommodate cyclic network

relations resemble a Bayesian network: a directed acyclic graph (DAG) with discrete random variables as vertices and conditional utility mass functions as edges. A DAG defines a model of preferential influences, which produces a unique set of *ex post* categorical utilities.[12] The game can be solved at this stage via standard solution concepts such as Nash equilibrium.

One selling point of conditional game theory is that it can give a clear operational meaning to the idea of an emergent preference ordering for a group, distinct from the preference orderings of the group members and based on the aggregation theorem (Stirling and Felin, 2013). They interpret such a *concordant utility*—the product of e's conditional utility and m's categorical utility, in our example—as a 'representation of the social consistency of the group, in that it provides a measure of the degree of severity of controversy.' (Stirling and Felin, 2013, 5). Hofmeyr and Ross (2019) contrast this way of approaching group-level preference to that of Bacharach et al. (2006). Let us now turn to this approach.

3.4 Models of team preference and team reasoning

Team preference and team-reasoning models (Sugden, 1993, 2000; Bacharach et al., 2006) were developed primarily to solve the puzzle of equilibrium selection in classical game theory. As noted above, despite the modeling of situations involving players with other-regarding preferences, many games remain uniquely unresolvable because there are multiple equilibria. The so-called Hi-Lo game highlights this problem after payoff transformations. It is a coordination game without conflicts of interest in which there are two pure-strategy Nash equilibria, (h_i, h_j) and (l_i, l_j), the former being clearly preferred by both players. In itself, the strategic rationality of best-reply reasoning—choose the action that maximizes one's own expected utility, given that others are doing the same—explains neither the observation that by far the majority of subjects choose (h_i, h_j) in experiments, thereby succeeding in more profitable coordination, nor the fact that most people intuit that (h_i, h_j) is the rational solution. Instead, individual strategic rationality needs to be supplemented with some auxiliary assumption, such as that the Pareto-dominating profile is a focal point that is either perceptually or normatively salient. According to Bacharach et al. (2006), this reveals a grave limitation of best-reply reasoning as the main theoretical assumption from which to derive solutions. Bacharach et al. (2006) propose that people instead apply *team-reasoning*: people (i) identify themselves as members of a team (group identification or we-framing), (ii) identify an action profile that is best for the team, (iii) identify their part in (ii), and (iv) play their

structures.

[12] Just as Fershtman and Segal (2018) restrict the range of influences within the observable others, Stirling and Felin (2013) introduce the idea of *sociation* to limit the number of the conjecture profiles that the agent considers.

part to achieve the team's goal. This process includes agency transformation, in contrast to the usual payoff transformation. In the Hi-Lo game, if both i and j identify with a team that consists of the two players, the process of team reasoning straightforwardly leads to (h_i, h_j) as the unique solution. Similarly, in the case of the employee and the manager, if e and m identify themselves with a company unit consisting of the two of them, an efficient scenario in which e exerts high working effort without m's high monitoring effort, that is, (h_e, l_m), can be achieved.

Bacharach et al. (2006) found support for this theory in the literature on group identification in the field of social psychology and on group-level selection in evolutionary biology. However, they do not use these theories simply to motivate or constrain the choice of a particular model of other-regarding preferences.[13] What is distinctive about their work is that the theory yields a hypothesis concerning the process through which the configuration of individual preference orderings endogenously activates group identification and team reasoning. Specifically, the *interdependence hypothesis* states that *ceteris paribus* the salience of three features of a game makes it more likely that the players adopt team reasoning. The first feature is *common interest*, defined as the fact that players prefer outcome o' (e.g.,(h_i, h_j)) to outcome o (e.g.,(l_i, l_j)); the second feature is *co-power*, the fact that o' can be brought about only by an appropriate combination of their actions; and the third is the existence of a Nash equilibrium that realizes o rather than o', making the attainment of o' *unassured* by their individualistic decision-making. The scope of the theory goes beyond pure coordination games such as Hi-Lo because these features are shared in mixed-motive coordination games and social-dilemma games,for example.

The hypothesis that individual preferences endogenously give rise to preference (and agency) transformations captures the fact that players' preferences are interdependent, as well as the fact that it is operationally meaningful in the framework of game theory to talk about group-level preferences. Similarly, in conditional game theory a network of individual categorical and conditional preferences gives rise to a group-level utility (concordant utility), from which the *ex post* preferences of each player are extracted (through *marginalization*). A major difference, however, is that conditional game theory uses classical solution concepts such as Nash equilibrium to derive equilibria of the game after social influence has been exerted through a preferential network, whereas models of team reasoning directly reject individual best-reply reasoning, thereby rejecting the basic logic underlying these solution concepts. This is more radical than the *refinements* of solution concepts such as fairness equilibrium. Given their radical nature, models of team reasoning remains a relatively unpopular approach among economists, although there is a sizable body of literature, both empirical (Smerilli, 2012; Guala et al., 2013) and philosophical (Gold

[13]On the use of the idea of group identity in economic modeling, see Chen and Li (2009). This is a variant of social-preferences model reviewed in Section 3.1.

and Sugden, 2007; Guala, 2017a).[14]

4 Methodological comparisons

Having presented the four main approaches to modeling social influences in economics, I now turn to some of their underlying methodological issues. The discussion in the following two subsections roughly covers epistemic and practical issues, respectively.

4.1 The domain of the models of socially influenced behavior

First of all, how do these models relate to each other? In terms of the mode of explanations, the models we have reviewed all seem to be located in the bottom-right area of Table 1, namely economic explanations of aggregate-scale phenomena on a relatively short time scale. The explananda of the models also largely overlap. For example, they all target stylized findings from non-cooperative experimental games as the main data against which their models are tested: cooperation in the social-dilemma game, coordination in mixed-motive or pure coordination games, fair offers in the ultimatum game, the dictator game, the trust game, and so on. Even the less empirically tested models, such as conditional game theory, explicitly use them as explanatory paradigm games (Hofmeyr and Ross, 2019; Stirling and Felin, 2013). Thus, they seem to compete in an empirical horse race in which not all of them hold true.[15] However, as we have seen in the quote from Camerer (2003) in Section 3.1 above, the model convergence is not necessarily assumed. Of course, this is unsurprising given the standard Kuhnian analysis of normal science: trade-offs between different epistemic values may not be uniquely solvable. However, I'd like to highlight a more specific factor here, which is how different models relate to the other types of explanations, such as the psychological (proximate) and the evolutionary (ultimate).

[14] On the philosophical level the meaning of group agency has been discussed in the context of social ontology. One major focus on the experimental level has been on whether social influences are exerted through changes in beliefs or changes in preference (e.g. Ellingsen et al., 2012). This question is rooted in social psychology, which concerns the mechanisms of group identification within the minimal-group paradigm (Jin and Yamagishi, 1997). Guala (2016) refers to team reasoning as a type of *simulation thinking*: a special case of individualistic reasoning. This line of research rejects best-reply as a model of interactive reasoning, without rejecting individualism. See also Guala (2018).

[15] Of course, if people are heterogeneous more than one model of social preference in a particular population can hold true at the same time, as explicitly assumed in the models of social norms I have discussed.

How should economic explanations of socially influenced behavior relate to the other types of explanations? First, it seems reasonable to demand that economic models should not *contradict* or be incompatible with them. In other words, the requirement of compatibility puts constraints on economic models. For example, models of social preference replace the default self-regarding model with various psychological hypotheses about what people care about, such as altruism, spite, social comparison, and intentional reciprocity (mostly based on the theorists' intuitive psychology). Models of team reasoning draws more from the literature on group identity in the field of social psychology. On the evolutionary front, Gintis's (2017) model of social norms, as well as Bacharach's (2006) model of team reasoning, explicitly consider evolutionary explanations such as group-level selection and gene-culture co-evolution as narrowing the range of preferences with which people are plausibly endowed. From this perspective, one could evaluate these models in terms of how well they cohere with explanations in other domains (interdisciplinary coherence).

However, things are not so simple because economists can have different understandings about what is distinct about the domain of economic explanations, vis-à-vis the other domains. I have already identified two points, a short time-scale (relative to evolutionary and historical processes) and aggregate-level phenomena (relative to individual psychological processes). Regarding the former, evolutionary game theory (which abstracts from human choices and models the replicator dynamics of strategies) implicitly moves toward a longer timescale, thereby competing less with other models.[16] Regarding the latter, all the models discussed in Section 3 are models of aggregate phenomena in the sense that they focus on the solutions of games among people with beliefs and preferences, categorical or conditional. However, different models seem to interpret this aggregation process differently. In the view of Bacharach et al. (2006), for example, the logic of best reply is a psychological hypothesis about human inference. This interpretation opens up the possibility to propose alternatives, such as team reasoning and solution thinking that give rise to aggregate-level phenomena. On the other hand, Hofmeyr and Ross (2019) and Ross and Stirling (2020) explicitly (and Fershtman and Segal (2018) implicitly) reject such a psychological interpretation of the strategic rationality of game theory. Commenting on Bacharach et al. (2006), Hofmeyr and Ross (2019) note:

> Game theory, like economics, is concerned with choices. If choice is *defined* in terms of outputs of reasoning processes, it follows that an account of team agency must be an account of reasoning. [...] However, in our view, a general theory of an aspect of agency, particularly eco-

[16]Sometimes the evolutionary approaches produce a psychological hypothesis positing that people mistake one-shot interactions for repeated interactions in the ancestral life. Thus formulated, this can be tested in experimental games.

> nomic agency, should reflect the more deflationary account of choice [...] According to this deflationary view, a behavior is chosen just in case it is subject to influence by incentives, regardless of whether the causal channel that links incentives and behavior involves deliberation.

From this deflationary perspective, a general-purpose model compatible with a wide range of psychological mechanisms—such as conditional game theory—is preferred. Models of team reasoning, in contrast, imply interactions among typically symmetrical players who team-reason, which is a strong restriction on the domain of the model, as well as on the mechanisms underlying choices. Similarly, models of social preference committed to particular psychological mechanisms are to be avoided, not necessarily because they have more free parameters for curve-fitting (some, e.g. Binmore and Shaked (2010) criticize these models as *ad hoc* for this reason), but because they constrain economic modeling too much, unnecessarily narrowing down its domain of applicability.[17]

Bicchieri's (2006; 2017) theory of social norms is ambivalent in its commitment to psychological mechanisms. On the one hand, the key constructs are belief-dependent preferences, within the structure of belief-desire psychology, but on the other hand, Biccchieri points out that the model should not be interpreted as a literal description of real psychological processes. In particular, Bicchieri (2006, 56) emphasizes the automatic way in which semantic priming activates a schema (or a script) that provides the context with a structure comprising a set of actions, events, people and their roles. This hypothesis relies heavily on the psychology of *unmotivated* actions or efficient social cognition, implying that most norm-following behavior is automatic. To reconcile these two aspects, Guala (2017b) interprets Bicchieri's theory as a kind of dual account in which two distinct psychological processes are at work ensuring the continuation and resilience of norms. In contrast, Ross et al. (forthcoming) show that conditional game theory can formally capture Bicchieri's theory of social norms by abstracting from the specifics of psychological processes.

How well do these models capture the aggregation processes of *social influence*? The answer depends on the meaning of social. Models of social preference are the most naive in identifying 'social' with 'other-regarding,' while assuming some norm-reference point as exogenously and unproblematically given. Models of social norms are more developed in explicitly referring to expectations shared by individuals and in modeling how they function as focal points or correlated equilibria.

[17]Being non-committal with respect to the psychology of choice does not mean that other disciplines play no role in debates in economics. In fact, Hofmeyr and Ross (2019) argue against team reasoning, drawing on evidence from developmental psychology to the effect that human infants (and possibly other primates) with limited reasoning capacities have a natural propensity to engage in joint actions. In this case, psychology is invoked not to constrain economic models, but to motivate keeping them more flexible.

These approaches model social influence as a network of expectations among variably conformist individuals. In contrast, models of conditional and team preferences hypothesize about endogenous processes through which configurations of individual preferences give rise to aggregate-level preferences, which in turn affect individual choices. Social influence is modeled as a network of preference change, with more or less specification of psychological mechanisms (models of team reasoning focuses on group identification, whereas conditional game theory remains neutral). With the exception of social preference models, the 'social' refers to aggregate dynamics distinct from individual characteristics, and in this sense those models constructively address the criticism that *homo economicus* is asocial, albeit in different ways.

4.2 Applications of the models of socially influenced behavior

The discussion in this subection concerns two issues related to the practical use of these models of socially influenced behavior, which is related but distinct from the issue of explanatory domains above.

A natural question in this regard is whether and how these models can be used to facilitate *social change*. For example, models of social norms suggest that rapid social change is possible through intervention in people's expectations, which are more malleable than the psychological traits such as conformist preferences, which are evolutionarily and developmentally more stable. A good example is the use of Bicchieri's model of social norms to demolish bad norms. Female genital cutting (FGC), which is a severe violation of human rights, is prevalent in many African countries. However, there is wide cross-country variety in the level of support for the practice among women who are responsible for having their daughters' genitals cut. The prevalence rate of FGC in Mali, for example, is around 85 percent, and the practice is supported by 76 percent of women aged between 15 and 49. In contrast, the level of support in Sudan, where the practice is even more prevalent (close to 90%), is less than 25 percent (Bicchieri, 2017, 46, Table 1.3). The implication is that FGC in Sudan is an unpopular norm sustained by pluralistic ignorance: the majority's private distaste for the practice is not observable, whereas its prevalence convinces those in the majority that there are high expectations of conformity. Thus, although FGC in Sudan might disappear relatively quickly if there were interventions to make those private preferences public, the practice in Mali will not change unless preferences are targeted more directly.

Undesirable equilibria include not only severe violations of human rights, but also socially tolerated behavior with substantial negative externalities, such as the unsustainable consumption of private goods or the violation of social distancing during a pandemic. Although the initial literature on *nudging* predominantly focused on pro-

moting individual well-being, drawing on behavioral models of individual decision-making, the prospect of nudging pro-social behavior, such as pro-environmental action, has recently been the subject of lively discussion in interdisciplinary journals (e.g., Schubert, 2017; Davis et al., 2018; Centola et al., 2018; Maki et al., 2019; Hagmann et al., 2019; Capraro et al., 2019; Otto et al., 2020; Kristal and Whillans, 2020). Models of socially influenced behavior could contribute to this effort by providing policymakers with recipes for behavioral change. If these models are expected to identify the right buttons to press for systemic and sustainable behavioral change, they have to correctly identify the key mechanisms that drive change and sustain desirable states, on both the individual-psychological and aggregate-social levels. Models of team reasoning have not been used as systematically as models of social norms, but they have a similar potential for application to behavioral change, particularly in contexts in which group identity or strong interdependence is salient (Nagatsu, 2015). Models of conditional preference have not yet been applied, but they could be useful if they could be operationalized and measured, and were to intervene on the level of *sociation*, that is the subset of others whose preferences matter in terms of the range of conjecture or observable behavioral preferences.

Models of social preferences are also used to plan a specific type of social change through the *mechanism design* of marketplaces. Bolton and Ockenfels (2012), for instance, considered the optimal design of online marketplaces. They found that buyers and sellers in an online second-hand commodity market showed a preference for reciprocal favor when writing reviews. In other words, a typical user writes an inflationary positive review about the other party anticipating reciprocation. The result is reciprocal equilibrium in which positive reviews dominate the platform. One way of evaluating the situation is to suggest that the welfare of the participants improves as a result of their reciprocal preference satisfaction, as a matter of definition or as a matter of feel-good psychology. However, Bolton and Ockenfels (2012) consider this to be a suboptimal situation that behavioral economic engineering should modify, because the reciprocal behavior decreases the quality of information about the reliability of the market participants, ultimately hurting them all and deterring potential participants from entering.[18] In contrast, in the design of the non-market-based voluntary provision of public goods, or common resource management, Gintis's model of cooperation for example suggests the importance of a critical proportion of strong reciprocators to ensure the efficient provision of public goods.

This comparison leads to another practical question: how should models of social interaction be used for welfare evaluation? In general, policy interventions

[18] *Airbnb.com*, for example, addresses this problem by forcing hosts and guests to review each other in a simultaneous-move game, in which you cannot see the other party's reviews on you until you have submitted your own.

in collective social behavior cannot beg the question of welfare judgement. The assumption in traditional welfare economics is that standard preferences capture everything economic agents care about, whatever they may be, which should be respected. It is suggested in the above case, on the other hand, that economists have more of a system-level approach in asking whether social preferences deserve the policy-makers' respect. They should be dampened sometimes, to safeguard efficient information flow in the marketplace, but they should be exploited other times, to facilitate the efficient provision of public goods.

Models of social preference as such seem unable to give a principled answer to the ethical questions about how much and which social preferences deserve respect. First, on the empirical level it seems difficult to find a universal social-utility function that accommodates all the data from different experimental games. Second and more fundamentally, models of social norms and conditional preferences imply that observed preferences are path-dependent and interdependent in many social contexts. In other words, such preferences are endogenously created in networks of social influences in a given environment. One cannot base welfare analysis of a system on social preferences if the latter is a function of the former.

An analogy with models of individual judgement and decision-making may be useful here. In the context of evaluating Prospect Theory as a behavioral alternative to the received Expected Utility Theory (EUT), Ross (2014) argues that the former will not supplant the latter. This assessment is based on the reasoning that there is no such thing as an ultimate theory of bounded rationality—be it Prospect Theory or some new alternative—because the economic rational agency with transitive and consistent preferences that EUT models is in itself emergent, scaffolded by the market by means of institutional and technological artefactual arrangements. Similarly, one could argue that it is impossible to find *the* ultimate theory of social preferences because preferences—be they self-regarding or other-regarding—are emergent in particular historical and institutional contexts. There is evidence from cross-cultural studies, for example, that preferences for cooperative and fair play in a range of two-person experimental games strongly correlate with the level of market integration in a given society (Henrich, 2004), and that preferences for hierarchical group dominance and institutionalized group-level discrimination (e.g., racism and sexism) are interdependent (Kunst et al., 2017). Sociologists have long pointed out that economic rational agency (with a self-interested calculus) is a product of the market-centered modern society, but the models of conditional preferences suggest that preferences are constructed *in general*: preferences for fairness and discrimination also seem to be a product of sociality.

If this analogy holds, in other words, if social arrangements cause more or less stable social preferences, how should one go about making welfare evaluations of social arrangements or institutional designs? Yet, none of the models of socially

influenced behavior reviewed in this chapter in themselves provide an unambiguous ethical framework. The literature on the ethical permissibility of social nudging based on these models cannot address the question either, because it takes asocial, autonomous individuals as the starting point of normative analysis. A pragmatic alternative is to judge what is reasonable case-by-case, focusing mostly on the local goals of specific marketplaces, such as online auction and matching platforms. However, some theorists have addressed the general normative question concerning the desirability of market economies (e.g. Gui and Sugden, 2005; Bowles, 2016; Ross, 2013).

5 Conclusion

In this chapter I have reviewed four distinct approaches to the economic modeling of socially influenced behavior, namely in terms of social preferences, social norms, conditional preferences, and team reasoning. I have put particular emphasis on their distinct methodological approaches to modeling social influence. In particular, I showed that these approaches differ not only in empirical and technical detail, but also in the interpretation of *sociality* in economic explanations. Some models (of social preference and team reasoning) interpret human sociality as biological or psychological constraints on the characteristics of preferences or the process of preference formation, whereas others (models of conditional preference and to some extent models of social norms) interpret it as contingent and interactive processes of the formation of aggregate expectations and preferences. As such, the latter approach addresses the asocial criticism of *homo economicus* in a more fundamental way than the former. Put differently, in the framework of Calcott's table (Table 1), the latter is conscious about the distinct features of economic explanations as non-reductive (as opposed to psychological proximate explanations) and constructive (as opposed to biological ultimate explanations). In this sense, economic explanations largely overlap with what we understand as sociological ones, although the former is usually associated with social engineering and management, whereas the latter with social critique.

Where should philosophers of economics go from these observations? Let me conclude by briefly speculating on two possible directions. One is a practical turn, involving broadening the methodological analysis of economic experimentation and modeling to incorporate their wider social (and economic!) conditions and aspirations. In particular, the narrow focus of methodologists on the epistemic success of models and experiments and their extrapolation could be widened to include assessment of the more constructive use of experimental and modeling tools in applied contexts (e.g., Redpath et al., 2018). The second is a normative turn, which seems necessary to properly evaluate the welfare implications of social arrangements which

themselves affect social preferences. This could be called the ethics of institutional design, but standard normative ethics, including the utilitarianism underlying welfare economics, is insufficient to the extent that it assumes stable preferences as given prior to social influence. Instead, something like what Slaby (2016) calls a *political philosophy of mind*—the reflective and systematic assessment of social arrangements—seems inevitable once the domain of economics is understood as distinctly social.

References

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal*, 100(401):pp. 464–477.

Bacharach, M., Gold, N., and Sugden, R. (2006). *Beyond individual choice: teams and frames in game theory*. Princeton University Press, Princeton, N.J.

Becker, G. S. (1974). A theory of social interactions. *The Journal of Political Economy*, 82(6):1063–1093.

Bicchieri, C. (2006). *The Grammar of Society*. Cambridge University Press, Cambridge, England.

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

Binmore, K. and Shaked, A. (2010). Experimental economics: Where next? *Journal of Economic Behavior & Organization*, 73(1):87–100.

Bolton, G. E. and Ockenfels, A. (2012). Behavioral economic engineering. *Journal of Economic Psychology*, 33(3):665–676.

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

Calcott, B. (2013). Why the proximate-ultimate distinction is misleading, and why it matters for understanding the evolution of cooperation. In Sterelny, K., Joyce, R., Calcott, B., and Fraser, B., editors, *Cooperation and its evolution*, chapter 13, pages 249–263. MIT press, Cambridge, MA.

Camerer, C. F. (2003). *Behavioral Game Theory*. Princeton University Press, Princeton, NJ.

Capraro, V., Jagfeld, G., Klein, R., Mul, M., and de Pol, I. v. (2019). Increasing altruistic and cooperative behaviour with simple moral nudges. *Scientific Reports*, 9(1):11880.

Centola, D., Becker, J., Brackbill, D., and Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119.

Chen, Y. and Li, S. X. (2009). Group identity and social preferences. *American Economic Review*, 99(1):431–57.

Davis, T., Hennes, E. P., and Raymond, L. (2018). Cultural evolution of normative motivations for sustainable behaviour. *Nature Sustainability*, 1(5):218–224.

Dhami, S. (2016). *The Foundations of Behavioral Economic Analysis*. Oxford University Press, Oxford.

Ellingsen, T., Johannesson, M., Mollerstrom, J., and Munkhammar, S. (2012). Social framing effects: Preferences or beliefs? *Games and Economic Behavior*, 76(1):117–130.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does Fairness Prevent Market Clearing? An Experimental Investigation*. *The Quarterly Journal of Economics*, 108(2):437–459.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fershtman, C. and Segal, U. (2018). Preferences and social influence. *American Economic Journal: Microeconomics*, 10(3):124–42.

Gintis, H. (2009). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton University Press, Princeton.

Gintis, H. (2017). *Individuality and Entanglement: The moral and material bases of social life*. Princeton University Press, Princeton, NJ.

Gold, N. and Sugden, R. (2007). Collective intentions and team agency. *The Journal of Philosophy*, 104(3):109–137.

Guala, F. (2012). Reciprocity: Weak or strong? what punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35.

Guala, F. (2016). *Understanding institutions: The science and philosophy of living together*. Princeton University Press, Princeton.

Guala, F. (2017a). Preferences: neither behavioural nor mental. *DEMM Working Paper*, Number 5.

Guala, F. (2017b). Review of cristina bicchieri's norms in the wild: how to diagnose, measure, and change social norms. oxford: Oxford university press, 2017, xviii + 239pp. *Erasmus Journal for Philosophy and Economics*, 10(1):101–111.

Guala, F. (2018). Coordination, team reasoning, and solution thinking. *Revue D'Économie Politique*, 128(3):355–372.

Guala, F., Mittone, L., and Ploner, M. (2013). Group membership, team preferences, and expectations. *Journal of Economic Behavior and Organization*, 86:183–190.

Gui, B. and Sugden, R., editors (2005). *Economics and Social Interaction: Accounting for Interpersonal Relations*. Cambridge University Press, Cambridge, England.

Gurven, M. (2004). Does market exposure affect economic game behavior? In Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., and Gintis, H., editors, *Foundations of human sociality: economic experiments and ethnographic evidence from fifteen small-scale societies*, chapter 7, pages 194–231. Oxford University Press, New York.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388.

Hagmann, D., Ho, E. H., and Loewenstein, G. (2019). Nudging out support for a carbon tax. *Nature Climate Change*, 9(6):484–489.

Henrich, J. P. (2004). *Foundations of human sociality: economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford University Press, Oxford.

Hofmeyr, A. and Ross, D. (2019). Team agency and conditional games. In Nagatsu, M. and Ruzzene, A., editors, *Contemporary philosophy and social science: An interdisciplinary dialogue*, chapter 3, pages 67–91. Bloomsbury Publishing, London.

Hsiung, B. (2001). A methodological comparison of Ronald Coase and Gary Becker. *American Law and Economics Review*, 3(1):186–198.

Jin, N. and Yamagishi, T. (1997). Group heuristics in social dilemma. *Shakai Shinrigaku Kenkyu*, 12(3):190–198.

Kahneman, D., Knetsch, J. L., and Thaler, R. H. (1986). Fairness and the assumptions of economics. *The Journal of Business*, 59(4):pp. S285–S300.

Kristal, A. S. and Whillans, A. V. (2020). What we can learn from five naturalistic field experiments that failed to shift commuter behaviour. *Nature Human Behaviour*, 4(2):169–176.

Kunst, J. R., Fischer, R., Sidanius, J., and Thomsen, L. (2017). Preferences for group dominance track and mediate the effects of macro-level social inequality and violence across societies. *Proceedings of the National Academy of Sciences*, 114(21):5407–5412.

Maki, A., Carrico, A. R., Raimi, K. T., Truelove, H. B., Araujo, B., and Yeung, K. L. (2019). Meta-analysis of pro-environmental behaviour spillover. *Nature Sustainability*, 2(4):307–315.

Mason, W. A., Conrey, F. R., and Smith, E. R. (2007). Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and Social Psychology Review*, 11(3):279–300. PMID: 18453465.

Mayr, E. (1961). Cause and effect in biology. *Science*, 134(3489):1501–1506.

Nagatsu, M. (2015). Social nudges: Their mechanisms and justification. *Review of Philosophy and Psychology*, 6:481–494.

Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockström, J., Allerberger, F., McCaffrey, M., Doe, S. S. P., Lenferna, A., Morán, N., van Vuuren, D. P., and Schellnhuber, H. J. (2020). Social tipping dynamics for stabilizing earth's climate by 2050. *Proceedings of the National Academy of Sciences*, 117(5):2354–2365.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 83(5):pp. 1281–1302.

Redpath, S. M., Keane, A., Andrén, H., Baynham-Herd, Z., Bunnefeld, N., Duthie, A. B., Frank, J., Garcia, C. A., Månsson, J., Nilsson, L., Pollard, C. R. J., Rakotonarivo, O. S., Salk, C. F., and Travers, H. (2018). Games as tools to address conservation conflicts. *Trends in Ecology & Evolution*, 33(6):415–426.

Ross, D. (2013). The evolution of individualistic norms. In Sterelny, K., Joyce, R., Calcott, B., and Fraser, B., editors, *Cooperation and its evolution*, chapter 1, pages 17–44. The MIT Press, Cambridge, MA.

Ross, D. (2014). *Philosophy of Economics*. Palgrave Macmillan, New York.

Ross, D. and Stirling, W. (2020). Economics, social neuroscience, and mindreading. In Harbecke, J. and Herrmann-Pillath, C., editors, *Social neuroeconomics: Mecchanistic integration of the neurosciences and the social sciences*. Routledge.

Ross, D., Stirling, W. C., and Tummolini, L. (forthcoming). Strategic theory of norms for empirical applications in political science and political economy. In Kincaid, H. and van Bouwel, J., editors, *Oxford Handbook of Philosophy of Empirical Political Science*. Oxford University Press, Oxford.

Schelling, T. C. (1960). *The strategy of conflict*. Harvard University Press.

Schubert, C. (2017). Green nudges: Do they work? are they ethical? *Ecological Economics*, 132:329 – 342.

Slaby, J. (2016). Mind invasion: Situated affectivity and the corporate life hack. *Frontiers in Psychology*, 7:266.

Smerilli, A. (2012). We-thinking and vacillation between frames: filling a gap in bacharach's theory. *Theory and Decision*, 73(4):539–560.

Stirling, W. C. (2012). *Theory of Conditional Games*. Cambridge University Press, Cambridge, UK.

Stirling, W. C. and Felin, T. (2013). Game theory, conditional preferences, and social influence. *PLoS One*, 8(2):e56751.

Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy*, 10(1):69–89.

Sugden, R. (2000). Team preferences. *Economics and Philosophy*, 16(2):175–204.

Tigar, M. E. and Levy, M. R. (1977). *Law and the Rise of Capitalism*. New York Monthly Review Press, New York.