Modeling norm-governed communities with conditional games: Sociological game-determination and economic equilibria

Don Ross*
School of Society, Politics, and Ethics, University College Cork
School of Economics, University of Cape Town
Center for Economic Analysis of Risk, Georgia State University

Wynn C. Stirling
Department of Electrical and Computer Engineering, Brigham Young University

Luca Tummolini
Institute of Cognitive Sciences and Technologies, Italian National Research Council
Institute for Future Studies, Sweden

Declaration of interest: None

Publication declaration: This paper has been made available as a Working Paper on the website of our research institute, the Center for the Economic Analysis of Risk, Georgia State University, but has not been submitted for publication elsewhere.

*Corresponding author: Don Ross
don.ross931@gmail.com
Address: Dun Rua, Woodhill Park, Tivoli, Cork, T23 TV7D, Ireland.

Modeling norm-governed communities with conditional games: Sociological game-determination and economic equilibria

Abstract

Most social scientists agree that informal norms constrain available equilibria in most human interactions. However, they do not agree on how to model them: economists often make them derivative of individual preferences, while a broader tradition in social theory understands them as exogenous social facts. Non-cooperative game theory more naturally accommodates the economists' approach. However, attention is increasingly attracted to recent work by economists who appreciate that the broader understanding may be important for full empirical adequacy. We focus on how game theorists might track this emerging shift. Extending Stirling's previously developed Conditional Game Theory, we model macrostructural processes of norm evolution through social influence diffusion in a way that relies on no exotic solution concepts, which in turn allows norms as social facts and norms as expressions of preferences to be modeled as evaluable complements, by analogy to the complementarity of cooperative and non-cooperative game solutions under the Nash program. The result can be understood as a way of specifying mutual constraints between economic models in which normative attitudes are exogenous, and sociological models that represent such attitudes as endogenous under power relationships and ontologies of social roles.

JEL codes: A12, A13, C70, D84, D85, D91

Keywords: social norms; social facts (sociology); conditional games; public goods game

1 Introduction

The role of informal norms in constraining available equilibria in social dynamics has been a subject of steadily increasing interest among social scientists. There is as yet no general theory of norms to play the role of a 'default model' against which contributions can be assessed. This is partly because there is not a consensus on whether it is better to model a general phenomenon of norm responsiveness, with specific normative contents treated as varying with circumstances, or whether emphasis should instead be given to exploring the functionality of specific norms - for, e.g., fairness, reciprocity, modesty, patriotism, etc. - as separate exercises, with general normativity being treated as an emergent construct. The latter approach is typically referred to as the 'social preferences' approach, and as such has been criticised for lack of desirable generality (Binmore 2010; Smith and Wilson 2019). Among efforts to model norm responsiveness in general, some influential models from economics are those of Sugden (1998), Bernheim (1994), Brock and Durlauf (2001), and Michaeli and Spiro (2015, 2017), which have inspired some experimental applications (Andreoni and Bernheim 2009; Andreoni, Nikiforakis, and Siegenthaler 2017).

The present essay is about general responsiveness to norms. Its intended contribution is methodological: we aim to show how a specific modeling technology – conditional game theory (CGT) – can be used to represent relationships between the stabilisation of norms in groups, and strategic choices of individual agents. The latter kind of process is typically modelled by economists, using standard game theory. The former has less clear-cut disciplinary 'ownership'. Norms arise and spread partly through cultural evolution, for which there is a rich literature linking formal anthropological models descended from Boyd and Richarson (1985) with evolutionary game theory (for example, Gintis 2009, 2016). On shorter-run scales, norms are promoted by exercises of interpersonal influence. Where such influence shapes people's conceptions of their identities, often mediated through occupation of professional and other recognised social roles, economists typically yield scientific responsibility to sociologists. The main explanation for this is only partly a matter of disciplinary inertia. It also stems from the fact economists are most at home building models in which agents' preferences are exogenously given and stable, whereas identity formulation, at least as standardly conceived, involves endogenous preference change. In making the obvious point that these various processes influence one another, we intend no challenge to historically useful divisions of interdisciplinary labour. However, as a matter of empirical fact, the processes obviously constrain and causally influence one another. We understand our work here as a way of focusing formally on these linkages. Thus we see it as contributing to cross-disciplinary unification, without any implication that disciplines should merge or fuse.

We revisit this general theme in the concluding discussion. We do so not through reflections on abstract philosophy of social science, but by reference to technical considerations that are our primary focus. These are illustrated by the core exhibit of the paper, a sample simulation that compares our approach to a prior contribution to the literature on modeling strategic effects of norms. This should be understood as a formal proof-of-concept, not as a substantive general result. We do not here present an intended empirical theory of the dynamics of norms. Rather, we furnish a modelling approach that we hold to be potentially useful for operationalising such theories in the lab or the field.

One feature of norms that is now emphasised in almost all models is their *conditionality*: agents' choices to conform their actions to norms are often sensitive to the extent to which they observe corresponding behaviour in those with whom they interact. Bicchieri (2006, 2017) distinguishes unconditional norms sharply from conditional ('social') ones, following the Kantian tradition in regarding the former as the special domain of morality. For reasons explored by Binmore (1994, 1998, 2005), we are not persuaded that the distinction is sufficiently behaviorally stable to bear the weight that Bicchieri assigns to it. That is, we suspect that only a very rare and unusual kind of agent would go on following a norm if this was costly to her and she expected literally *no one* else to also follow it. This skepticism, about the empirical extent of morality in Kant's sense, only strengthens the idea that normative conditionality is central to understanding *choice-sensitive* sociality (as opposed to the hardwired sociality of ants, termites, and cells) and ethics.

There is a weak sense in which conditionality is sown into the very fabric of game theory: whether a given choice is an element of an equilibrium strategy for an agent is a function of what other agents choose. Therefore, whether an action in a strategic setting maximizes an agent's utility function in *ex ante* expectation depends on whether it is a best *reply* to other players' expected choices. But this only captures what Bicchieri (2006, 2017) calls 'empirical' expectations, expectations about what others will in fact do, and, in conditions of uncertainty about utility functions, what others expect others to do, and expect others to expect others to do, and so on recursively. These kinds of expectations are modeled using well developed game-theoretic solution concepts, including Bayes-Nash equilibrium (Harsanyi 1967) and quantal response equilibrium (McKelvey and Palfrey 1995; Goeree, Holt, and Palfrey 2016), and, for extensive-form games, sequential equilibrium (Kreps and Wilson 1982). Conditional responses can therefore, in principle, be purely implicit in best-response patterns across ranges of games.

The conception of norms of interest to us here, however, models them as exogenous social structures that agents encounter as elements of their environments. Agents can then have (conditional or unconditional) preferences and beliefs about whether they themselves, and other agents with whom they interact, regulate their choices, or *should* regulate their choices, by reference to them. This conception reflects the generally accepted empirical fact that norms display intertemporal and cross-generational persistence on a scale that is longer than individual preferences. Furthermore, as discussed and modelled by Kuran (1995), agents can under some circumstances go on following norms that are no longer preferred by majorities, or even by anyone at all, if something (e.g., fear of shame, or of the secret police) systematically interferes with revelation of true preferences.

On this 'social facts' (Gilbert 1989) conception of norms, we might suppose that they are subject to conditional preference in a special sense that goes beyond equilibrium dependence (though must be compatible with whatever definition of equilibrium features in the analysis). This is at the heart of Bicchieri's (2006, 2017) proposed special utility structure for application to games in which expectations about norms are relevant to choice. As reviewed in Ross, Stirling, and Tummolini (2023), Bicchieri and her various co-authors have, however, tended to fall back on formulations more consistent with the social preferences approach when analysing data from their own experimental lab.

Here we focus attention on models for application to choice data estimation that remain consistent with a ‘social facts’ conception of norms. Specifically, we attend to the utility structure proposed by Kimbrough and Vostroknutov (KV) (2016) for application to public goods games, dictator, ultimatum, and trust games run in the laboratory. This work represents a significant advance with respect to marrying generality and identifiability in choice data. We aim to deepen understanding of it as a model of empirical normative dynamics by embedding it in a more general framework.

For reasons we will indicate, KV’s model only accommodates norms for which social welfare increases in the number of followers. Nor does it allow for an agent to persistently follow a norm she would be better off abandoning. Thus it does not allow for representation of ‘Kuran cases’, that is, norms that survive due to strategic preference falsification and trap societies in inferior states relative to all plausible welfare criteria until something breaks informational symmetry. As the KV (2016) model is constructed to describe the specific class of social dilemma games indicated above, and is not intended to be general, it might be objected that in doing this we patch a tyre that isn’t losing any air. However, as we describe in the Discussion section of the paper, the subsequently developed general model of Kimbrough and Vostroknutov (2020a) also excludes Kuran cases.

Kuran cases remain out of reach of KV’s models for a deep reason. Although KV represent norms as social facts in the sense that agents have preferences about them *as* norms, norms so understood lack the kind of *causal force* that most social theorists regard as distinctive to genuine norms. As Kuran (1995) recognises, *persistent norms shape preferences*. A norm that endures for a time due to preference falsification may cause agents’ preferences to shift to accommodate it, and may consequently cease to be a barrier to Pareto improvement because it alters the location of the Pareto frontier, even while leaving market-valued wealth or income unimproved. That is, there are *social* adaptive preferences. Individual adaptive preferences are often regarded as phenomena that undermine standard normative economics. This is because they seem inconsistent with the idea of consumer sovereignty that in turn supports welfarism, the doctrine that the best economic policy is the one that most efficiently satisfies subjective preferences (Sen 1992). As we discuss in Section 4, the implications of social adaptive preferences for normative economics are more nuanced and arguably more interesting. As with individual adaptive preferences, they are sometimes defensive responses to oppression, of sub-cultures or sub-communities, but they are also the basis of often celebrated ‘expansions of the moral circle’ (Singer 1981).

Such endogenous preference change is generally excluded in standard game theory, though there are exceptions (e.g. Bisin and Verdier 2001). Providing a generalised approach to representing strategic preference adjustment is the primary objective of CGT (Stirling 2012, 2016). This is an extension of standard game theory that is designed to bring the social phenomenon of *mindshaping* (Zawidzki 2013) within the ambit of strategic choice. Mindshaping occurs when agents coordinate with one another by aligning ex ante uncertain preferences. One useful way to understand social norms is as outputs of mindshaping processes in networked groups.

KV’s models retain fixed preferences, and in most applications of interest to economists there are good reasons for this, as we will discuss. The model we develop is thus not intended to improve or replace either the KV2016 special model or the KV2020 general model of choice under norms. It is, rather, intended to represent a strategic process by which the normative expectations that the KV models treat as exogenous exert special, endogenous influence *as norms*. The games we model and those modelled by KV mutually constrain one another according to a relationship described by Ross (2005) under the label of ‘game determination’. As we explain, this can be understood, following Ross (2014), as step along a path to unifying macrostructural sociological and microeconomic theory in the formalism of game theory.

The paper is constructed as follows. In Section 2 we explain its motivations in more detail, and relate these to the motivations underlying the KV2016 special model of norm-governed social dilemma games. In Section 3 we compare simulations of play of a representative such game, the public goods game, using, respectively, KV's model based on application of standard game theory and our construction of a model using CGT. Section 4 discusses wider implications of this comparison. In particular, we sketch a conception of an account of norms in which macrosociological and microeconomic analyses play distinct but complementary roles. Their relationship is characterised as analogous to the Nash program for using cooperative and noncooperative game theory in tandem to exploit the strengths and avoid the limitations of each. Section 5 concludes and looks toward follow-up work.

2 Modelling norms with game theory: from mindreading to normative mindshaping

2.1 The importance of mindshaping processes for norm emergence

If norms are generally conditional, then agents playing games in norm-governed social contexts are under pressure to form expectations by making inferences about the normative attitudes of other players. As KV wrote in a 2013 Working Paper that was the ancestor of KV2016, "one way of thinking about social cognition is that an important part of 'theory of mind' is the ability to infer social norms from context". They here refer to the extensive literature on 'mindreading' (see Nichols and Stich 2003). Where conditional norms are in play, mindreading goes beyond updating of priors about other players' preferences based on observations that regard all of their their actions as symmetrically informative, because normative preferences will be revealed by choices only when their activating conditions happen to be satisfied.

Zawidzki (2013) argues that mindreading is, in general, much more difficult than theorists have tended to assume, given the evidence available to participants in the typical interactions that mindreading hypotheses are intended to explain. Zawidzki musters substantial evidence that most phenomena that other theorists have characterised as instances of mindreading are in fact manifestations of mind*shaping* processes, in which interacting parties exert influence on one another to generate alignment of mutually attributed preferences and beliefs. This interpretation of evidence is set within a wider *externalist* conception of the ontology of 'propositional attitudes' (PAs), such as belief and desire, that has become the dominant view among philosophers (Dennett 1987; McClamrock 1995; Clark 1997; Bogdan 1997). According to externalists, PAs are not private psychological states of individuals that must be inferred from behaviour, but publicly negotiated *interpretations* of relatively consistent *patterns* in behaviour (including verbal behaviour) that people use to make sense of one another and construct their own images of themselves as relatively coherent selves with unfolding biographies that comprise meaningful narratives. Mindshaping is typically relatively effortless and implicit because it is simply equivalent to the normal interpersonal construction of shared social reality.

Zawidzki does not deny that successful mindreading sometimes occurs. But it is necessary only in circumstances where a party to an interaction is thought to be concealing or misrepresenting their self-ascribed beliefs or preferences, perhaps for strategic reasons, or where there is asymmetry between parties in the sophistication with which contents of PAs are distinguished and articulated (e.g., a psychotherapist and a patient, a parent and a child, or a scientific analyst and a research subject). Zawidzki argues, based on review of experiments by cognitive scientists, that successful mindreading in such cases is far from assured; and that when mindreading *is* successful it depends on foundations of prior mindshaping with respect to background beliefs and preferences that provide contextual leverage for inference.

Mindshaping processes are not amenable to easy representation in standard game-theoretic models because they involve changes in preferences (as well as in beliefs). In standard game theory, preferences are the basic arguments for the utilities associated with outcomes, and in that sense define and individuate players. Since games with imperfect information typically involve strategic signaling and screening of information about which players hold relevant beliefs, or about types of players where types are distinguished by their preference profiles, it is natural to interpret games between people as models of mindreading rather than mindshaping.

One way of understanding mindshaping is as describing the psychological dynamics of *socialisation*. This could be socialisation of a child or adolescent person into normative adulthood, or of an immigrant into a community, or of a new employee into a corporate culture. These phenomena, because they involve adapting individuals to social roles, are typically regarded as the domain of sociologists though, again, they have occasionally been modelled by economists (Bisin and Verdier 2001; Akerlof and Kranton 2010). One might frame the division of labour between sociologists and economists, in an idealised unified social science, as follows: sociologists study the dynamics of agent formation and stabilisation, that is, mindshaping processes, which are preconditions for applications of economics to people who still face coordination challenges *given* maintained preferences and beliefs. Sometimes this coordination is brought about transparently by competitive markets or analysis of games of perfect information, and sometimes it is brought about by mindreading modeled as the calculation of Bayesian equilibria in extensive-form games where some information sets contain nodes >1. (Of course, microsociologists in the tradition of Goffman [1959] also study specific, short-run interactions using different methods; our comments here concern one possible view of microeconomics from a sympathetic sociologist's perspective, not imagined 'essences' of disciplines.)

This idealised division of labour blurs where the modeling of response to norms is concerned. After developing their special model of norms for ultimatum, dictator, public goods, and trust games, KV (2016) write that "[t]he most important unanswered question ... and the one that we hope this research will encourage others to ask, is 'where do norms come from?'" (p. 635). Their subsequent general model addresses this question in *one* sense: according to that model, different norms arise for different strategic contexts, but always as solutions to a general problem of minimising aggregate dissatisfaction with actual outcomes by comparison with achievable counterfactual alternatives. Binmore (1994, 1998), on the other hand, extensively addresses the question in a more fundamental sense, appealing to evolutionary psychology to explain why people are sensitive to norms in the first place. And we might turn to evolutionary game theory to account for general features of human sensitivity to normative influence. In abstracting away from idiosyncratic preferences of individual agents, explanations that rely on evolutionary modeling like Binmore's are relatively long-run models while KV's models, in which agents with fixed preferences maximise utility, are short-run ones. Though there is not yet a standard account of the ways in which the long-run models constrain the short-run models, there is no shortage of formal work relating evolutionary and classical game-theoretic solution concepts (e.g. Weibull 1995), or of reflection on the nature of their consilience (Binmore 1994, 1998; Gintis 2009, 2016).

According to Ross (2005, 2006, 2008), however, a crucial middle layer of analysis between evolutionary and standard game-theoretic analysis is needed but has been largely neglected in the literature. Early evolutionary models of norm emergence were inspired by biology and abstracted away from individual agents to focus on the adaptive value of behavioural strategies viewed as traits (Sugden 1986/2004; Binmore 1994; but see Young 2015 for a different approach). On the other hand, in standard game-theoretic models agents appear fully formed, with stable utility functions (see also Davis 2010). As discussed above, however, individuals are socialised into relative stability with respect to preferences (and beliefs) by mindshaping. Furthermore, mindshaping has a strategic dimension in which already socialised agents play essential roles, and less-than-fully socialised 'patients' co-develop their own agency by strategically blending accommodation, resistance, and creativity. Ross was therefore motivated to seek modeling approaches that, unlike biologically-inspired evolutionary models, preserve agency, while also allowing for representation of preferences that shift under social pressure.

2.2 Socialisation as game determination between generations

The basic device that Ross (2006, 2008) exploits to try to achieve this ambition is an overlapping generations model, lifted from its traditional macroeconomic setting (Samuelson 1958) and applied in a microeconomic, game-theoretic environment. He models what he calls 'game determination' using non-terminating sequences of extensive-form games across generations of agents (in which generations can, depending on the intended scale of analysis, include different life stages of a single biological individual). Each sequential triplet of games has the following structure. The first stage, the 'determining' game G'_1, involves players who adhere to norms, understood following Binmore (1994, 1998) as equilibrium selection devices, along with 'pre-socialised' players whose utility functions are generated by replicator dynamics that represent natural selection (including Baldwin effects) of human psychology. The outcomes of determining games are rules (extensive-form structures and strategy sets) of the 'determined' game stage G_2 that follows. All players' utility functions in all stages incorporate, to an exogenously parameterised degree, interest in the welfare of their descendants over a horizon of two subsequent periods. Pre-socialised players implicitly choose norms these successors will be endogenously motivated to follow given the structure of the determined game G_2 that is selected from among the equilibria of the determining game G'_1. In G_2 the pre-socialised agents are replaced by socialised successors, who might or might not have modified the norms of their 'tutors'. Determined games, which involve only socialised players, represent bargaining over distributions of resources that generate relative bargaining power in the subsequent determining game G'_3. In G'_3 agents who played G'_1 as socialised agents have died, replaced by pre-socialised descendants. This pattern iterates indefinitely.

Ross intended this game determination model to be a more or less literal implementation of the account given by Binmore (1994, 1998). It was thus acknowledged to be incomplete in the same respects that Binmore's philosophical theory is incomplete. First, it allows no scope for altruism about the interests of non-descendants, so coalitions of interests can only arise implicitly in the underlying replicator dynamics. But clearly humans often choose to participate in coalitions. An implication of this is that, as in the special model of KV2016, there is no scope for representation of norms that prescribe differential treatment of in-group and out-group members. Third, normative influence, i.e. mindshaping, is exclusively intergenerational, and there is none between generational peers. Fourth, the model is not general, in that the modeler can freely choose the parametric structures specifying interest in descendants' welfare; and the parametric structures governing these interests as between socialised and pre-socialised agents must be chosen independently because socialised agents live for two periods and pre-socialised agents live for only one period.

The game determination model's retention of a core restriction of standard game theory, that no agent's utility function changes, is also the main source of the limitations identified above. These limitations are removed in CGT.

2.3 Socialisation as cyclic influence in social networks

The primary motivation for CGT is the insight that, just as agents adjust uncertain beliefs on the basis of observing beliefs of others, so they may resolve ambivalent preferences by comparing them with preferences of others in cultural or commercial or political reference groups. CGT is characterised by Ross and Stirling (2021) as a formal theory of mindshaping. The core technical manoeuvre in the construction of the theory is to apply the syntax of epistemic probability to the practical domain, i.e., to incentivised choice. It is therefore equipped to model what might be regarded as cognitive dynamics using data compatible with revealed preference theory.

There are three stages involved in representing and solving a conditional game: socialisation, diffusion, and deduction. Socialisation is achieved by expressing preferences via *conditional payoffs* that reflect modulation by agents of their utility structures as functions of the preferences of those who socially influence them. This form of conditional reasoning is formally analogous to the use of conditional probabilities to modulate beliefs as functions of statistical influence. Just as a set of statistically dependent random variables can be expressed as a directed graph (a Bayesian network) with random variables as the vertices and conditional probabilities as the edges, so a community of socially influenced agents can be expressed as a directed graph (a *social influence network*) with agents as the vertices and conditional payoffs as the edges. Then, by formal analogy to the way in which statistical dependency is diffused throughout a Bayesian network to create the joint probability of the random variables as the product of the individual conditional probabilities, social influence is diffused throughout a social influence network to create a *coordination function* as the product of the individual conditional payoffs. In epistemology individual probabilities of the random variables are deduced from the joint probability by marginalisation; so, analogously, in CGT socially influenced payoffs of the individual agents are deduced from the coordination function by marginalisation. Following marginalisation, games are solved by application of standard equilibrium concepts for normal-form games (e.g., Nash equilibrium, quantal response equilibrium). CGT is equivalent to standard noncooperative game theory if no social influence exists (i.e., a social influence network has no edges), which might be because no agents are *ex ante* ambivalent about their preferences, and the conditional payoffs thus are identical to categorical payoffs.

As originally developed in Stirling (2012) and Tummolini and Stirling (2020), conditional game theory was restricted to acyclic networks, thereby confining the theory to hierarchical networks in which influence flows are unidirectional. However, Ross and Stirling (2021) extend the theory to account for cyclic influence by applying Markov chain convergence theory. Ross, Stirling and Tummolini (2023) further extend the theory to incorporate choice under uncertainty (following a specification of Prelec 1998) and maximisation by agents of rank-dependent utility (Quiggin 1982).

The reader can consult a more formal outline of the core features of CGT in Appendix A.

Philosophically, we view the relationship between game determination theory (GDT) and CGT as follows. We are convinced that, as a matter of empirical fact, processes of socialisation and diffusion set the conditions for human interactions that game theorists model and solve deductively. However the mechanisms that transmit social influence are modelled mathematically, we expect Ross's (2005) philosophical account to apply to them. But for now we reserve 'GDT' for what is actually on the table by way of real theory, namely, Ross's specific (2006, 2008) overlapping-generations model. Then we can say that both GDT and CGT aim to represent mindshaping. Because GDT preserves a one-to-one mapping of agents to utility functions, it is more conservative in its way of extending standard noncooperative game theory. But the price of this is high: it can only represent intergenerational mindshaping. CGT, in also capturing peer-to-peer mindshaping, is more general. This also comes with a cost: CGT cannot usefully be applied to extensive-form games in which solutions rely on identification of outcomes with fixed preferences. Of course a normal-form game is formally a set of extensive-form games. When a normal-form game with categorical preferences is 'relaxed' in CGT to allow for conditional preferences, the associated set of extensive-form games typically explodes. As we argue in Section 4, this apparent cost reflects the scale shift between macrostructural and microeconomic modeling - it is why macrostructural modeling is insufficient for identifying the phenomena that interest microeconomists. Following arguments in Ross (2014), we believe that a general philosophy of social science should aim to show how disciplines can make complementary contributions while remaining distinct in the formal analyses specified by their general theories. The kind of complementarity we have in mind is exemplified in the application in Section 3 below. But it is *merely* exemplified. The project of theoretically specifying general mechanisms by which processes of socialisation and deductions of solutions to strategic interactions of fully socialised agents constrain one another awaits future work. We will return to sketch the challenge less cryptically at the end of the paper.

2.4 Normative mindshaping: from categorical to conditional norm-dependent utility

With this context in place, we return to KV's modeling strategy. Building on previous work (Kessler and Leider 2012), KV assume that the human tendency to comply with social norms can be fruitfully described with a *norm-dependent* utility function like the following:

$$u_i(x) = x - \phi_i|\eta - x| \tag{1}$$

Here x corresponds to the agent's material payoff, η is the action that is most socially appropriate in a given context (the norm), $\phi_i \geqslant 0$ is an individual parameter that specifies the individual sensitivity to norms, and the distance $|\eta - x|$ between the socially appropriate action and the actual one captures the disutility from norm violation. Notice that, if defined in this way, norm-dependent utility does not amount to a kind of social preference (in the technical sense), since the norm-following agent only cares about the extent to which the considered action conforms to the norm η and not about the payoff other agents derive. A social preference model instead typically assumes that a norm is embodied in a *social* utility where a "taste for fairness" (or some other specific social value) modulates how much an individual agent values alternative outcomes (e.g. Fehr and Schmidt 1999). In contrast with this approach, an agent who is simply sensitive to norms in general can consistently be "selfish" on some occasions and "generous" on others if the norm that is relevant demands different actions in different contexts. Consider, for instance, that the same norm that specifies a "fair" distribution of a windfall gain may also prescribe a "selfish" one when resources have instead been earned (Oxoby and Spraggon 2008).

Notice that this formulation of the norm-dependent utility does not theoretically constrain where the norm η comes from. Thus, especially for the purposes of experimental work, KV rely on the experimental task proposed by Krupka and Weber (2013) in which, by exploiting the incentive structure of a pure coordination game, the relevant norm underlying a given social situation is inferred by eliciting the action which is believed to be the most appropriate by a reference group of other people (i.e. second-order normative beliefs or normative expectations in the sense of Bicchieri 2006). Finally, complementary to this procedure, KV introduce an additional task to also estimate, albeit indirectly, ϕ_i, conceived as a general, idiosyncratic trait revealing a *categorical desire* to actually comply with normative expectations.

To better understand how the KV model works consider, for instance, a standard Public Goods (PG) game. As the workhorse of the experimental study of cooperation, the PG is commonly proposed as a model of a situation in which a group has the opportunity to invest in a common project with the potential to benefit all of its members. This collective welfare-optimizing outcome however risks not to be achieved because the temptation to free-ride on the effort of others and the fear of being unfairly exploited may induce widespread defection. Indeed, the decline of cooperation in repeated Public Goods games is one of the most robust and best replicated findings in experimental economics (Ledyard 1995; Chaudhury 2011). Starting with the seminal contributions by Yamagishi (1986) and Fehr and Gachter (2000), it has been repeatedly observed that existence of a punishment mechanism in game rules can indeed sustain cooperation in these difficult contexts. In real informal settings amongst people, the relevant mechanisms are often naturally interpreted as norms (Smith and Wilson 2019). Consistently with this interpretation, KV has shown that groups composed of norm followers (i.e. members with high ϕ) are indeed capable of sustaining cooperation at similar levels even if punishment is not available, which is something that is not possible for groups of norm-breakers (i.e. with members with low ϕ) that were unable to resist the well-known pattern of cooperation decline.

To rationalize these results, KV propose that the relevant norm in a repeated PG is a norm of *conditional* cooperation. In a repeated PG such a norm encourages the ith player, denoted X_i, to contribute η at the start of the interaction (e.g. in period 1) and to keep contributing similarly in the next rounds *if* the other members have contributed the appropriate amount too. Violation of the norm by others make X_i's violation appropriate. Indeed, KV have shown that, under such a norm, cooperation can be sustained as a Perfect Bayesian Nash Equilibrium if players' norm-sensitivity parameters ϕ_i are sufficiently high as well as being *believed to be sufficiently high* by all. In other words, in keeping with standard philosophical interpretation of game theory applied to humans, norms influence behaviour via a mind-reading process by which norm-followers infer the action that others believe to be most appropriate as well as their reciprocal desires to act according to it.

Adopting the mindshaping perspective, however, opens up a complementary possibility. Instead of defining a (categorical) norm-dependent preference to cooperate conditionally on the cooperation of others, i.e. a preference for reciprocity in action, we can begin by specifying a *conditional* norm-dependent preference towards cooperation as such, i.e. the norm emerges from convergence in the preferences for cooperation in a group. In this approach norm-dependent preferences form due to the fact that the utility that a norm follower derives from cooperating in a PG is affected by the possibility that cooperation is assumed to be the action that one's reference group prefer the most too. Viewed in this way, conditional norm-dependent preferences are the product of a mind-shaping, interactive influencing dynamics that aim to create and stabilize behavioral patterns in interactive contexts.

As we explain in Section 4, we frame the technical work of the next two sections as follows. KV, in both their special and general models, provide an improved microeconomic analysis of the operation of norms. Such analysis applies under the idealisation of fixed utility functions. For the huge range of applications that matter to economists, who specialise in understanding the marginal effects of changes in incentives, typically under institutional constraints that are commonly known, this idealisation is powerful, the basis of elegant deductive solutions. But on less granular scales, the scales on which social dynamics forge utility functions through mindshaping, the idealisation must be relaxed.

The work in the next two sections below should be understood as exemplifying this relaxation. The example is not intended to show the full potential for representing social structure using CGT, which would constructing sub-networks with heterogeneous agents. Such construction is illustrated in Ross, Stirling, and Tummolini (2023). The example here is offered as a formal proof-of-concept with the purpose of demonstrating the causal effects of preference diffusion, that is, normative mindshaping, within a single community whose network connections are assumed to be symmetrical throughout.

3 Simulating a Public Goods Game Using Conditional Game Theory

3.1 Unconditional KV Utility Model

We consider the public goods game scenario introduced by KV (2016), which involves a collective $\{X_1, \ldots, X_n\}$ of players where the players possess a common action set $\mathcal{A} = \{x_1, \ldots, x_N\}$ and each player has the option of contributing part of her endowment to the community and receiving some fraction of the total contribution in return. The KV payoff model for agent X_i is

$$\pi_i(a_i, a_{-i}) = u_{i,i}(a_i) + u_{i,-i}(a_i, a_{-i}) \tag{2}$$

where $a_i \in \mathcal{A}$ is an action by X_i and $a_{-i} = \{a_1, \ldots, a_n\} \backslash \{a_i\}$ is the set of acts of all agents $X_{-i} = \{X_1, \ldots, X_n\} \backslash \{X_i\}$, $u_{i,i}(a_i)$ is the part of X_i's payoff that she improves by choosing a_i, and $u_{i,-i}(a_i, a_{-i})$ is the part of the payoff chosen for X_i by X_{-i}.

In the KV (2016) model set-up, the payoffs are defined over the unit interval. In order to conduct simulations, we need to replace the unit-interval actions set employed by KV with a finite action set. We consider a set of four agents $\{X_i, i = 1, \ldots, 4\}$ with common action $\mathcal{A} = \{x_1, x_2, x_3\}$ with $x_1 > x_2 > x_3$ expressed in units of dollars. For our model we set $\mathcal{A} = \{50, 25, 0\}$. The resulting payoff function is

$$\pi_i(a_i, a_{-i}) = x_1 - a_i + \alpha \sum_{j=1:n} a_i \,, \tag{3}$$

where $u_{i,i}(a_i) = x_1 - (1-\alpha)a_i$ and $u_{i,-i}(a_i, a_{-i}) = \alpha \sum_{j \neq i} a_j$ with $\alpha \in [0,1]$. KV (2016) extends this model to create a repeated-play game defined over a sequence of stages $t = 0, 1, 2, \ldots$ that includes a *norm-response* component $\phi_i\, \rho(s_{t-1})\, g\big(\|\eta - a_{it}\|\big)$, yielding, for $t = 1, \ldots$,

$$\pi_i(a_{it}, a_{-it}) = x_1 - a_{it} + \alpha \sum_{j=1:n} a_{jt} - \phi_i\, \rho(s_{t-1})\, g\big(\|\eta - a_{it}\|\big)\,, \tag{4}$$

where

- $\phi_i \geqslant 0$ is a dimensionless parameter indicating the sensitivity of player X_i to deviations from the norm η;
- $\rho(s)$ is the norm response function of the average contribution $s_t = \frac{1}{4}\sum_{i=1:4} a^*_{it}$ of all players, defined as
$$\rho(s_t) = \begin{cases} 1 & \text{if } s_t = \eta \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$
where a^*_{it} is X_i's choice at stage t;
- g is a strictly convex increasing function that represents the disutility of deviating from the norm and is defined as
$$g(z) = x_1 \frac{1 - e^{\frac{z}{x_1}}}{1-e}\,, \tag{6}$$
with $\|\eta - a_i\| = |\eta - a|$, the absolute value of the difference (notice that g is scaled to units of dollars), yielding $g\big(\|\eta - x_1\|\big) = g\big(\|\eta - x_3\|\big) = g(\eta) = 18.877$ and $g\big(\|\eta - x_2\|\big) = 0$;
- $\eta = \$25$ is the norm that governs agents' behavior.

KV proposes the following strategy, as presented in KV(2016) Appendix B.1. At stage $t = 0$, each player takes an action a^*_{i0} drawn from a distribution F that is common knowledge. Since the initial actions strongly influence subsequent behavior, it is imperative that these actions are randomized such that the average s_0 is uniformly distributed. For each $s \in \{\$0, \$25, \$50, \$75, \$100, \$125, \$150, \$175, \$200\}$, Table 1 lists the combinations of $a_1 + a_2 + a_3 + a_4 = s$ for $(a_1, a_2, a_3, a_4) \in \boldsymbol{\mathcal{A}}$. To ensure that the actions drawn from these subsets are uniformly distributed, we must draw from the set of constant-sum subsets with probability proportioned to the number of elements in the subset, and then draw uniformly from that subset. For example, there are 10 ways to achieve a sum of \$150 (i.e., $301 \cup 220$) so this subset is selected with probability $10/81$, and an element is drawn from this set with probability $1/10$. We also require that each element of the set $\{\phi_1, \phi_2, \phi_3, \phi_4\}$ be drawn via the independent and identically uniform distribution over the interval $[\phi^* - \epsilon,\ \phi^* + \epsilon]$, where $\epsilon > 0$ and ϕ^* is the threshold value such that the maximizing choice for payoff (4) switches from $a^*_i = x_3$ to $a^*_i = x_2$ when $\phi_i \geqslant \phi^*$ (the value for ϕ^* will be computed subsequently). Table 2 lists the cumulative distribution function for constant-sum partitions for each contribution sum and the corresponding inverse distribution. The randomization of the initial action proceeds as follows: Let a be drawn from a uniform distribution over $[0, 1]$, apply $F^{-1}(a)$ to identify the constant-sum partition, and draw the initial actions $(a_{10}, a_{20}, a_{30}, a_{40})$ from a uniform distribution over the selected constant-sum partition.

KV's payoff model for $t > 0$ is

$$a^*_{it} = \arg\max_{a \in \mathcal{A}} \{-(1-\alpha)\, a - \phi_i\, \rho(s_{t-1})\, g(\|\eta - a\|)\}\,. \tag{7}$$

Table 1: Initial action partition sets for constant sum.

x_1	x_2	x_3	# Partitions	Sum
4	0	0	1	$200
3	1	0	4	$175
3	0	1	4	$150
2	2	0	6	$150
2	1	1	12	$125
2	0	2	6	$100
1	3	0	4	$125
1	2	1	12	$100
1	1	2	12	$75
1	0	3	4	$50
0	4	0	1	$100
0	3	1	4	$75
0	2	2	6	$50
0	1	3	4	$25
0	0	4	1	$0

Table 2: Cumulative distribution and inverse distribution for choice initiation.

x	$x \leqslant 0$	$x \leqslant 25$	$x \leqslant 50$	$x \leqslant 75$	$x \leqslant 100$	$x \leqslant 125$	$\leqslant 150$	$x \leqslant 175$	$x \leqslant 200$
$F(x)$	$\frac{1}{81}$	$\frac{5}{81}$	$\frac{15}{81}$	$\frac{31}{81}$	$\frac{50}{81}$	$\frac{66}{81}$	$\frac{76}{81}$	$\frac{80}{81}$	1
a	$a \leqslant \frac{1}{81}$	$a \leqslant \frac{4}{81}$	$a \leqslant \frac{15}{81}$	$a \leqslant \frac{31}{81}$	$a \leqslant \frac{50}{81}$	$a \leqslant \frac{66}{81}$	$a \leqslant \frac{76}{81}$	$a \leqslant \frac{80}{81}$	$a \leqslant 1$
$F^{-1}(a)$	0	25	50	75	100	125	150	175	200

Let

$$\begin{aligned}
\pi_i(x_1) &= \begin{cases} -(1-\alpha)x_1 - \phi_i\, g(\|\eta - x_1\|) = -(1-\alpha)50 - 18.877\,\phi & \text{if } s_{t-1} = \eta \\ -(1-\alpha)x_1 = -(1-\alpha)50 & \text{otherwise} \end{cases} \\
\pi_i(x_2) &= \begin{cases} -(1-\alpha)x_2 - \phi_i\, g(\|\eta - x_2\|) = -(1-\alpha)25 & \text{if } s_{t-1} = \eta \\ -(1-\alpha)x_2 = -(1-\alpha)25 & \text{otherwise} \end{cases} \\
\pi_i(x_3) &= \begin{cases} -(1-\alpha)x_3 - \phi_i\, g(\|\eta - x_3\|) = -18.877\,\phi_i & \text{if } s_{t-1} = \eta \\ -(1-\alpha)x_3 = 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{8}$$

and the optimal choice is

$$\max\{p_i(x_1), p_i(x_2), p_i(x_3)\} = \begin{cases} -(1-\alpha)25 & \text{if } \phi_i > \phi^* \text{ and } s_{t-1} = \eta \\ -18.877 & \text{otherwise,} \end{cases} \tag{9}$$

where

$$\phi^* = \frac{(1-\alpha)25}{18.877}\,, \tag{10}$$

resulting in

$$a^*_{it} = \begin{cases} x_2 & \text{if } \phi_i > \phi^* \text{ and } s_{t-1} = \eta \\ x_3 & \text{otherwise.} \end{cases} \tag{11}$$

Specifying behavior in this repeated-play game is straightforward. Each player is initialized with a randomly selected action. If the average of their initial actions is consistent with expectations under the collective welfare improving norm and the sensitivity parameter ϕ_i is greater than ϕ^* for all players, then they all conform to the norm and continue to do so at subsequent stages. However, if the average of the initial actions is not consistent with normative expectations, then each player seeks to maximize her payoff at subsequent stages, regardless of the existence of a norm. However, even if the average initial action is the norm-consistent, if $\phi_i < \phi^*$ for any X_i, then she will ignore the norm, thereby causing the other players to do likewise.

3.2 Norm-Compliant Utility Model

As defined by KV (2016), the only mechanism to induce compliance with the norm is the history provided by s_{t-1}. The function ρ signals to the network at stage t that the agents at stage $t-1$ did or did not comply with the norm and invokes a penalty if they did not, but that is not the only mechanism for exertion of social pressure. We will suppose that if a norm exists then the players are aware of it, and that this awareness is crucial for motivating punishment and given that knowledge, each player may, regardless of knowledge of the past, feel social pressure to comply with the norm. The purpose of our simulation is to demonstrate the effectiveness of this supposition in the CGT framework.

We model social influence influence via conditional utility. The utilities of standard game theory involve deductive inferences, that is, the expression $u(a) > u(a')$ establishes that a is preferred to a'. Deductive inferences are also familiar from probability theory. Let $pr[\cdot]$ denote a subjective probability mass function. The expression $pr[a] > pr[a']$ establishes that a is more likely than a'. But probability theory also supports a different kind of inference; namely conditional inferences. The conditional probability statement $pr[a|b] > p[a'|b]$ *does not* provide sufficient information to conclude that a is more likely than a' without a supporting statement regarding the likelihood of b. Similarly, we may adopt the syntax of probability theory to define a *conditional utility* of the form $u(a|b) > u(a'|b)$, meaning that if b is actualized, then a is preferred to a'.

Following this line of reasoning, let us define the conditional utility for the public goods game. Let $\mathbf{\Sigma}$ denote the set of all possible summative arrangements for $X_{-i} = \{X_1, \ldots, X_4\} \backslash \{X_i\}$, the subset of agents excluding X_i. Ordered lexicographically, the elements of $\mathbf{\Sigma}$ are

$$\begin{aligned} \Sigma_{111} &= x_1 + x_1 + x_1 \\ \Sigma_{112} &= x_1 + x_1 + x_2 \\ \Sigma_{113} &= x_1 + x_1 + x_3 \\ \Sigma_{121} &= x_1 + x_2 + x_1 \\ &\vdots \\ \Sigma_{133} &= x_1 + x_3 + x_3 \\ \Sigma_{233} &= x_2 + x_3 + x_3 \\ \Sigma_{333} &= x_3 + x_3 + x_3 \,. \end{aligned} \tag{12}$$

For any profile $[a_{it}, a_{jt}, a_{kt}, a_{lt}]$, we define the *exclusion sum*

$$\sigma_{-it} = a_{jt} + a_{kt} + a_{lt} \,, \tag{13}$$

as the sum of conjectures of all players *excluding* a_{it}, the conjectured contribution of X_i. Thus, σ_{-it} constitutes a conditioning conjecture set by X_i for X_{-i}. There are seven possible summative values for σ_{-it}, comprising the set

$$S = \{0, 25, 50, 75, 100, 125, 150\}\,. \tag{14}$$

Let

$$\mathcal{S}_r = \big\{(a_{j_r,q}, a_{k_r,q}, a_{l_r,q}) \in \mathcal{A}^3\colon a_{j_r,q} + a_{k_r,q} + a_{l_r,q} = s_r\big\} \tag{15}$$

denote the set of conditioning subprofiles whose sum equals s_r for $s_r \in S$ and $q = 1, \ldots, N_r$, where N_r is the number of subprofiles in $\mathcal{S}_r$. Table 3 displays the sets of all possible combinations, expressed in lexicographical order, of the ways $\mathcal{A}^3$ can be partitioned into constant s_r subsets.

Table 3: Conditioning conjecture partition sets for constant s_r.

x_1	x_2	x_3	# Partitions	Sum
3	0	0	1	\$150
2	1	0	3	\$125
2	0	1	3	\$100
1	2	0	3	\$100
1	1	1	6	\$75
1	0	2	3	\$50
0	3	0	1	\$75
0	2	1	3	\$50
0	1	2	3	\$25
0	0	3	1	\$0

Given the norm η, the **norm-compliance partition**, denoted $\mathcal{S}_\eta$, is the subset of $\mathbf{\Sigma}$ whose elements sum to 3η. For $\eta = \$25$,

$$\begin{aligned}\mathcal{S}_\eta = 030 \cup 111 = \big\{&(x_1, x_2, x_3), (x_1, x_3, x_2), (x_2, x_1, x_3), (x_2, x_2, x_2),\\ &(x_2, x_3, x_1), (x_3, x_1, x_2), (x_3, x_2, x_1)\big\}\,.\end{aligned} \tag{16}$$

A conjecture σ_{-it} is **norm-compliant** if $\sigma_{-it} \in \mathcal{S}_\eta$. The norm-compliant payoff function is

$$\pi_{i|-i}(a_{it}|\sigma_{-it}, s_{t-1}) = \begin{cases} x_1 - a_{it} + \alpha(a_{it} + \sigma_{-it}) - \phi_i(a_{it}, s_{t-1})\big[g\big(\|\eta - a_{it}\|\big) - x_1\big] & \text{if} \quad \sigma_{-it} \in \mathcal{S}_\eta \\ x_1 - a_{it} + \alpha(a_{it} + \sigma_{-it}) & \text{if} \quad \sigma_{-it} \notin \mathcal{S}_\eta \end{cases} \tag{17}$$

where, for $t = 0$, $s_{t-1} = \varnothing$. Thus, if X_i conjectures that X_{-i} will be norm compliant and she is favorably disposed toward the norm, then she would also favor others complying. If she conjectures that X_{-i} will not be norm compliant (i.e., she conjectures that X_{-i} favors outcomes in $\{x_2\}^C$), then she will seek, unconstrained, to maximize her utility without regard for the existence of a norm. The norm, then, conditions actions in the sense of Bicchieri (2006, 2017). X_i's norm-compliance sensitivity, ϕ_i, is now a function of her conjectured action a_{it} and the history s_{t-1}.

3.3 Two-Agent Conditional Public Goods Game

To define a conditional game we must transform $\pi_{i|-i}$ to become a mass function, yielding

$$u_{i|-i}(a_{it}|\sigma_{-it}, s_{t-1}) = \frac{\pi_{i|-i}(a_{it}|\sigma_{-it}, s_{t-1})}{\sum_{q=1:4} \pi_{i|-i}(a_{qt}|\sigma_{-it}, s_{t-1})} \,. \tag{18}$$

In the interest of clarity (and brevity) we first develop a one-shot public goods conditional game model for only two agents; extending it to four agents is then conceptually straightforward.

The most significant difference between this model and the KV model is the presence of direct linkages between the two agents whereby each agent modulates her preferences as a function of conjectures regarding the preferences of the other. Furthermore, these preference relationships are reciprocal: X_1 exerts influence on X_2, who influences X_1, and so forth, thus, creating a cyclic network of the form

$$\begin{tikzcd}
& s_{t-1} \arrow[dl] \arrow[dr] & \\
X_1(t) \arrow[rr, bend left] & & X_2(t) \arrow[ll, bend left]
\end{tikzcd} \tag{19}$$

where the arrows indicate the direction of social influence from the influencer to the influencee and s_{t-1} represents the information available at time $t-1$ that influences the agents at time t according to the utility model $\{u_{1|2s}(a_1|a_2, s), u_{2|1s}(a_2|a_1, s), u_S(s)\}$. One of the consequences of expressing conditional utility with the syntax of probability theory is that the mathematical machinery that has been developed in the probability context can be imported into conditional game theory. In particular, the Markov chain convergence theorem may be applied to characterize this mindshaping. Let τ denote an iteration index which may be an interval, but may also be viewed as an iterated calculation index that represents an adjustment process, or tatonnement, as agents respond to their social environment. Suppose, before the social engagement begins (i.e., $\tau = 0$), that each agent attaches a conditional utility of the form $u_{i|s}(a_{it}|s_{t-1}; 0)$. Analogous to the way beliefs are combined via the chain rule according to the syntax of probability theory to create a joint model of belief, we apply this syntax to combining preferences to create a joint model of preference, yielding

$$w_{ij|s}(a_{it}, a_{jt}|s_{t-1}; \tau) = u_{i|js}(a_{it}|a_{jt}, s_{t-1}) w_{j|s}(a_j|s_{t-1}; \tau) \tag{20}$$

for $i, j \in \{1, 2\},\ i \neq j$, where $u_{i|js}(a_{it}|a_{jt}, s_{t-1})$ is defined by the problem statement.[1] The term $w_{j|s}(a_j|s_{t-1})$ is the initial condition for X_i's utility as conditioned only by s_{t-1}. Fortunately, as we will subsequently establish via the Markov chain convergence theorem, all initial conditions converge to the same steady-state value. At iteration $\tau > 0$, each agent may update her individual utility as conditioned on s_{t-1} via marginalization, yielding

$$w_{i|s}(a_{it}|s_{t-1}; \tau) = \sum_{a_{jt}} w_{ij|s}(a_{it}, a_{jt}|s_{t-1}; \tau) = \sum_{a_{jt}} u_{i|js}(a_{it}|a_{jt}, s_{t-1}) w_{j|s}(a_j|s_{t-1}; \tau - 1) \,. \tag{21}$$

Notice that we do not sum over s_{t-1} since the sum of contributions at stage $t-1$ is common knowledge at stage t.

Expressing this relationship using matrix theory notation yields

$$\mathbf{w}_{i|s}(\tau) = T_{i|js} \mathbf{w}_{j|s}(\tau - 1) \,, \tag{22}$$

[1] The general form of the chain rule expressed in terms of conditional probability mass functions is $p(x, y|z) = p(x|y, z)p(y|z)$.

where

$$\mathbf{w}_{i|s}(\tau) = \begin{bmatrix} w_{i|s}(x_1|s;\tau) \\ w_{i|s}(x_2|s;\tau) \\ w_{i|s}(x_3|s;\tau) \end{bmatrix} \tag{23}$$

and

$$T_{i|js} = \begin{bmatrix} u_{i|js}(x_1|x_1, s_{t-1}) & u_{i|js}(x_1|x_2, s_{t-1}) & u_{i|js}(x_1|x_3, s_{t-1}) \\ u_{i|js}(x_2|x_1, s_{t-1}) & u_{i|js}(x_2|x_2, s_{t-1}) & u_{i|js}(x_2|x_3, s_{t-1}) \\ u_{i|js}(x_3|x_1, s_{t-1}) & u_{i|js}(x_3|x_2, s_{t-1}) & u_{i|js}(x_3|x_3, s_{t-1}) \end{bmatrix} \tag{24}$$

is the transition matrix for the linkage $X_i \to X_j$ as conditioned by s_{t-1}. Applying this expression iteratively yields

$$\mathbf{w}_{i|s}(\tau) = T_{i|js}\mathbf{w}_{j|s}(\tau-1) = T_{i|js}T_{j|is}\mathbf{w}_{i|s}(\tau-2) = \cdots = (T_{i|js}T_{j|is})^{\tau}\mathbf{w}_{i|s}(0) \tag{25}$$

or, more compactly,

$$\mathbf{w}_{i|s}(\tau) = T_{i|s}^{\tau}\mathbf{w}_{i|s}(0)\,, \tag{26}$$

where $T_{i|s} = T_{i|js}T_{j|is}$ is the closed-loop transition matrix.

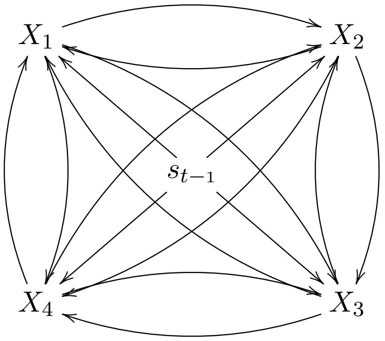
The Markov chain convergence theorem establishes that, if $T_{i|s}$ satisfies the regularity condition that all entries of T_{is}^{m} must be greater than zero for some integer m, then

$$\lim_{\tau\to\infty}\mathbf{w}_{i|s}(\tau) = \lim_{\tau\to\infty} T_{i|s}^{\tau}\mathbf{w}_{i|s}(0) = \overline{\mathbf{w}}_{i|s} \tag{27}$$

for all initial conditions $\mathbf{w}_{i|s}(0)$, where $\overline{\mathbf{w}}_{i|s}$ is the eigenvector corresponding to the unique unit eigenvalue of $T_{i|s}$. Thus, one does not need to actually perform the iteration; one can simply compute the eigenvector.

3.4 Four-Agent Conditional Public Goods Game

We now apply the above development to a four-agent conditional public goods game employing the conditional utilities defined in (18), yielding the graph



(28)

where the center node is s_{t-1}. The multidirectional influence paths greatly complicate the analysis of this game; thus we are motivated to develop a Markov equivalent graph of the form[2]

$$\begin{tikzcd}
X_4 & & X_3 \\
X_1X_2X_3 \arrow[u] \arrow[rr] & & X_4X_1X_2 \arrow[u] \arrow[dd] \\
 & s_{t-1} \arrow[ul] \arrow[ur] \arrow[dl] \arrow[dr] & \\
X_2X_3X_4 \arrow[uu] \arrow[d] & & X_3X_4X_1 \arrow[ll] \arrow[d] \\
X_1 & & X_2
\end{tikzcd} \tag{29}$$

where we have exchanged a network graph with single-agent vertices and multidirectional edges for a network graph with multi-agent vertices and unidirectional edges.[3] The issue now is to achieve Markov equivalence by defining edges that preserve the conditional relationships. Applying the chain rule, the transition from $X_jX_kX_l \to X_iX_jX_k$ conditioned on s_{t-1} is, suppressing the stage index t,

$$w_{ijk|s}(a_i, a_j, a_k|s) = \sum_{a'_j, a'_k, a'_l} w_{ijk|j'k'l's}(a_i, a_j, a_k|a'_j, a'_j, a'_k, s)\, w_{j'k'l'|s}(a'_j, a'_k, a'_l|s)\,, \tag{30}$$

where primes distinguish between the conditioning conjectures (a'_j, a'_j, a'_k) and the conditioned conjectures (a_i, a_j, a_k). Suppressing arguments and applying the chain rule,

$$w_{ijk|j'k'l's} = w_{k|ijj'k'l's}\, w_{ij|j'k'l's} \tag{31}$$

where, again applying the chain rule

$$w_{ij|j'k'l's} = w_{j|ij'k'l's}\, w_{i|j'k'l's} \tag{32}$$

yields

$$\begin{aligned} w_{ijk|j'k'l's}(a_i, a_j, a_k|a'_j, a'_k, a'_l, s) = w_{k|ijj'k'l's}(a_k|a_i, a_j, a'_j, a'_k, a'_l, s) \\ w_{j|ij'k'l's}(a_j|a_i, a'_j, a'_k, a'_l, s)\, w_{i|j'k'l's}(a_i|a'_j, a'_k, a'_l, s)\,. \end{aligned} \tag{33}$$

However, The conditional functions $w_{k|ijj'k'l's}$ and $w_{j|ij'k'l's}$ involve self-conditioning, that is, the conjectured actions for X_j and X_k appear as both conditioning actions (marked by primes) and conditioned actions. Thus, these mass functions are degenerate; hence,

$$\begin{aligned} w_{k|ijj'k'l's}(a_k|a_i, a_j, a'_j, a'_k, a'_l, s) &= \begin{cases} 1 & \text{if } a_k = a'_k \\ 0 & a_k \neq a'_k \end{cases} \\ w_{j|ij'k'l's}(a_j|a_i, a'_j, a'_k, a'_l, s) &= \begin{cases} 1 & \text{if } a_j = a'_j \\ 0 & a_j \neq a'_j\,, \end{cases} \end{aligned} \tag{34}$$

[2] An important property of network theory is that a graph of a network is *not* the network; rather, it is a representation of the network, and representations are not unique. An alternative graph of a network is said to be *Markov equivalent* if it preserves all of the conditional relationships.

[3] The graph displayed in (29) employs clockwise rotation; an equivalent graph using counterclockwise rotation is also possible.

and (33) becomes

$$w_{ijk|j'k'l's}(a_i, a_j, a_k|a'_j, a'_k, a'_l, s) = \begin{cases} w_{i|j'k'l's}(a_i|a'_j a'_k, a'_l, s) & \text{if } a_j = a'_j, a_k = a'_k \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

Markov equivalence will thus be assured if

$$w_{i|j'k'l's}(a_i|a'_j a'_k, a'_l, s) = u_{i|j'k'l's}(a_i|a'_j a'_k, a'_l, s)\,, \tag{36}$$

and it follows that (suppressing the stage index t),

$$w_{ijk|j'k'l's}(a_i, a_j, a_k|a'_j, a'_k, a'_l, s) = \begin{cases} u_{i|j'k'l's}(a_i|a'_j a'_k, a'_l, s) & \text{if } a_j = a'_j, a_k = a'_k \\ 0 & \text{otherwise} \end{cases} \tag{37}$$

for $i|jkl \in \{1|234, 2|341, 3|412, 4|123\}$.

Given these conditional linkages between the subgroups, we may close the loop by defining the subgroup-to-subgroup conditional transition matrices $T_{ijk|jkls}$ connecting subgroup vertex $X_j X_k X_l$ to $X_i X_j X_k$ given s_{t-1}, yielding

$$\mathbf{w}_{ijk|s} = T_{ijk|jkls}\mathbf{w}_{jkl|s}\,, \tag{38}$$

where

$$\mathbf{w}_{ijk|s} = \begin{bmatrix} w_{ijk|s}(x_1, x_1, x_1|s) \\ w_{ijk|s}(x_1, x_1, x_2|s) \\ w_{ijk|s}(x_1, x_1, x_3|s) \\ \vdots \\ w_{ijk|s}(x_1, x_3, x_3|s) \\ w_{ijk|s}(x_2, x_3, x_3|s) \\ w_{ijk|s}(x_3, x_3, x_3|s) \end{bmatrix}, \qquad \mathbf{w}_{jkl|s} = \begin{bmatrix} w_{jkl|s}(x_1, x_1, x_1|s) \\ w_{jkl|s}(x_1, x_1, x_2|s) \\ w_{jkl|s}(x_1, x_1, x_3|s) \\ \vdots \\ w_{jkl|s}(x_1, x_3, x_3|s) \\ w_{jkl|s}(x_2, x_3, x_3|s) \\ w_{jil|s}(x_3, x_3, x_3|s) \end{bmatrix}, \tag{39}$$

and

$$T_{ijk|jkls} = \begin{bmatrix} w_{ijk|jkls}[x_1, x_1, x_1|x_1, x_1, x_1, s] & \cdots & w_{ijk|jkls}[x_1, x_1, x_1|x_3, x_3, x_3, s] \\ w_{ijk|jkls}[x_1, x_1, x_2|x_1, x_1, x_1, s] & \cdots & w_{ijk|jkls}[x_1, x_1, x_2|x_3, x_3, x_3, s] \\ w_{ijk|jkls}[x_1, x_1, x_3|x_1, x_1, x_1, s] & \cdots & w_{ijk|jkls}[x_1, x_1, x_3|x_3, x_3, x_3, s] \\ \vdots & & \vdots \\ w_{ijk|jkls}[x_1, x_3, x_3|x_1, x_1, x_1, s] & \cdots & w_{ijk|jkls}[x_1, x_3, x_3|x_3, x_3, x_3, s] \\ w_{ijk|jkls}[x_2, x_3, x_3|x_1, x_1, x_1, s] & \cdots & w_{ijk|jkls}[x_2, x_3, x_3|x_3, x_3, x_3, s] \\ w_{ijk|jkls}[x_3, x_3, x_3|x_1, x_1, x_1, s] & \cdots & w_{ijk|jkls}[x_3, x_3, x_3|x_3, x_3, x_3, s] \end{bmatrix}. \tag{40}$$

The closed-loop transition matrices are computed as

$$T_{ijk|s} = T_{ijk|jkls}T_{jkl|klis}T_{kli|lijs}T_{lij|ijks}\,. \tag{41}$$

Once these closed-loop transition matrices are defined, we apply the Markov chain convergence theorem to compute the converged functions $\overline{\mathbf{w}}_{ijk|s}$. These converged coordination vectors are used to compute the individual converged utilities via

$$\overline{\mathbf{w}}_{i|s} = \begin{bmatrix} \overline{w}_{i|s}(x_1) \\ \overline{w}_{i|s}(x_2) \\ \overline{w}_{i|s}(x_3) \end{bmatrix} = T_{i|jkls}\overline{\mathbf{w}}_{jkl|s}\,, \tag{42}$$

where

$$T_{i|jkls} = \begin{bmatrix} u_{i|jils}(x_1|x_1,x_1,x_1,s) & \cdots & u_{i|jkls}(x_1|x_3,x_3,x_3,s) \\ u_{i|jils}(x_2|x_1,x_1,x_1,s) & \cdots & u_{i|jkls}(x_2|x_3,x_3,x_3,s) \\ u_{i|jils}(x_3|x_1,x_1,x_1,s) & \cdots & u_{i|jkls}(x_3|x_3,x_3,x_3,s) \end{bmatrix}. \tag{43}$$

The converged cyclic network is

$$\begin{tikzcd}
X_4 & & X_3 \\
X_1X_2X_3 \arrow[u, "T_{4|123s}"] \arrow[rr, squiggly, "T_{412|123s}"] & & X_4X_1X_2 \arrow[u, "T_{3|412s}"'] \arrow[dd, squiggly, "T_{341|412s}"] \\
 & s_{t-1} \arrow[ul] \arrow[ur] \arrow[dl] \arrow[dr] & \\
X_2X_3X_4 \arrow[uu, squiggly, "T_{123|234s}"] \arrow[d, "T_{1|234s}"'] & & X_3X_4X_1 \arrow[ll, squiggly, "T_{234|341s}"] \arrow[d, "T_{2|341s}"] \\
X_1 & & X_2
\end{tikzcd} \tag{44}$$

where the links defined by the symbol $\rightsquigarrow$ indicate that the network has reached steady-state once convergence is achieved.

3.5 Simulation Design

The simulation consists of two simulation experiments to compare two populations, both playing the repeated public goods game as defined in KV (2016). Each population comprises four agents $\{X_1, X_2, X_3, X_4\}$ engaging in K trials, with each trial comprising L stages. Population A plays the straight KV public goods game, and Population B plays the same game except for intermittent mindshaping episodes. This is a modeling convenience. We suppose that in reality mindshaping would occur continuously and incrementally. However, such continuous preference modulation cannot be directly represented consistently with standard game theory. We in effect probe the effects of background preference diffusion using comparative statics. The variable of interest is the frequency of norm-compliant choices between A populations and B populations.

3.5.1 Population A

The members of Population A are all independently initialized at stage $t = 0$ according to the distribution defined by Table 2, yielding initial random choices$(a_{10}, a_{20}, a_{30}, a_{40})$. The payoffs for stages $t > 0$ are given by (4), with $\alpha = 1/2$, $\eta = \$25$, and g defined by (6) . The threshold sensitivity level is ϕ^* as defined by (10), yielding $\phi^* = 0.662$. The sensitivity parameters ϕ_i for each trial are chosen independently according to the uniform distribution $U[\phi^* - \epsilon,\ \phi^* + \epsilon]$. The optimal solutions for $t > 0$ are defined by (11). Thus, if $\phi_i > \phi^*$ for all agents and $s_{t-1} = \eta$, then they all play the norm for all subsequent stages. However, if $\phi_i < \phi^*$ for any X_i, then she plays to maximize her payoff, which causes all agents to play selfishly for all subsequent stages. This is the critical feature that is explored via Population B. Namely, *if an agent comes to the social engagement with low norm-sensitivity, she may be susceptible to the influence of others*, and there may be reachable equilibria in which she is induced to adjust her norm sensitivity accordingly.

3.5.2 Population B

The members of Population B are initialized with a mindshaping exercise using the norm-compliant utility defined by (17). The distinctive feature of this utility is that, rather than being exogenously set ex ante, it is susceptible to modulation as a function of both the previous state s_{t-1} and σ_{-it}, the conditional conjectures of X_{-i}. (In fact, we may view the product $\phi\,\rho$ in (4) as a coarse modulation of ϕ as a function of s_{t-1}— an on-off switch.) Although we model ϕ_i as a function of both σ and s_{t-1}, in this section we focus on conditionalization via σ_{-it}. Given her conjecture that X_{-i} will follow the norm, if she were inclined to be a contributing citizen of the community, she would have an incentive to also follow the norm rather than risk the disapproval of her fellow citizens. This would motivate her to increase her norm sensitivity. On the other hand, if she were not normatively compatible with X_{-i}, she would not be so inclined.

Our simulation model for Population B proceeds as follows. Each agent is initialized with CGT using the norm-compliant payoff function defined by (17). In order to make comparisons with the basic KV model, we employ the same ϕ_i parameters as used by KV and we modulate them by multiplicative parameters as follows:

$$\hat{\phi}_i(a_{it}, s_{t-1}) = \begin{cases} \gamma(a_{it})\,\phi_i & \text{if } \sigma_{-it} \in \mathcal{S}_\eta \\ \phi_i & \text{otherwise,} \end{cases} \tag{45}$$

with $\gamma(a_{it}) = \gamma_q \geqslant 0$ if $a_{it} = x_q$, $q \in \{1, 2, 3\}$, where γ_2 increases norm compliance sensitivity by conjecturing increased norm sensitivity if X_{-i} conjectures norm compliance and γ_1 and γ_3 reduce norm compliance sensitivity if X_{-i} were to reject norm compliance.

Our simulation for Population B uses the outcome from the CGT first stage for subsequent stages using the basic KV model, which are interrupted at random stages by inserting the CGT model as a mindshaping reinforcement. The simulation consists of $K = 100$ trials with repeated games of length $L = 50$ stages. Table 4 displays the number of instances of norm compliance for both populations for several values of γ_i.

Table 4: Simulation results.

$(\gamma_1,\ \gamma_2,\ \gamma_3)$	Population A	Population B
(0.5, 2.5, 0.5)	194	392
(0.2, 2.5, 0.2)	185	2598
(0.7, 3.0, 0.7)	124	1568
(1.0, 3.0, 1.0)	105	147
(0.5, 3.0, 0.5)	126	4653
(0.5, 2.5, 0.2)	164	2303
(0.2, 2.5, 0.5)	152	368

Our simulations establish that mindshaping via CGT does indeed increase the frequency of norm compliance.Figure 1 displays simulation results as a function of γ_2 for various combinations of (γ_1, γ_3). These plots demonstrate consideraable mindshaping effectiveness sensitivity to γ_2, with significantly less sensitivity to γ_1, resulting in three clusters of graphs for $\gamma_2 \in \{0.2, 0.5, 0.7\}$. Mindshhaping effectiveness approaches zero as γ_2 declines toward 2.2 for any values of the other parameters. Figure 2 displays norm compliance frequency with (γ_1, γ_2) fixed at $(0.5, 2.3)$ as γ_3 increases from zero, which indicates that mindshaping becomes essentially ineffective for $\gamma_3 > 0.15$. We did not attempt a specification search for the equation governing the parameter relationships due to the arbitrary nature of the parameter restrictions; such theory would serve no generalising purpose. The point simply is to show that for some parameter values mindshaping makes a dramatic difference.

The threshold $\gamma_2\phi^*$ is significantly higher than the threshold $\phi^* = 0.662$ for the basic KV game due to the interrelationships that exist among the agents. In the B population, X_{-i} exerts direct social influence

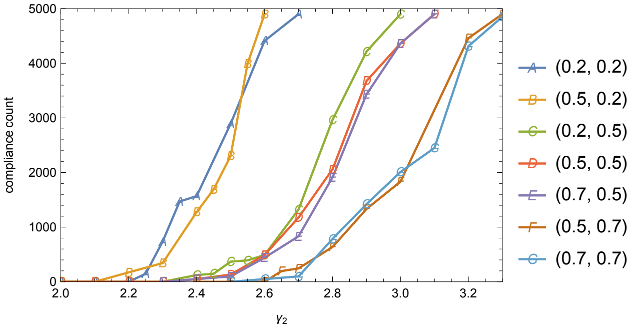


Figure 1: Norm compliance frequencies for selected parameter values

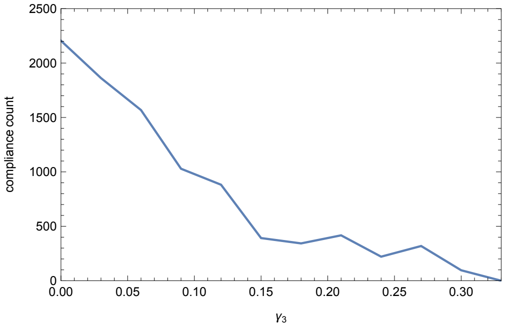


Figure 2: Norm compliance frequency versus γ_3 for $(\gamma_1, \gamma_2) = (0.5, 2.3)$.

on X_i, in contrast to the indirect social influence that the history from stage $t - 1$ exerts at stage t in the A population. Thus, although setting $\phi_i > \phi^*$ will certainly cause the payoff defined by (18) to favor x_2 over x_3 viewed in isolation from all other agents, when modeled as a part of the community where members are subjected to direct social influence, a more substantial increase in ϕ_i is required for x_2 to be favored over x_3 (see (17) and (45)).

These results establish that as the multiplier increases for x_2 and decreases for x_1 and x_3, the frequency of norm compliance increases. It is important to appreciate the distinction between the public goods game as defined by the basic KV model and the CGT-based public goods game. With the basic model, although each agent is influenced by the history of past actions, there are no explicitly defined social relationships among them. In particular, their norm sensitivities are held constant over all of their possible actions. By contrast, the CGT model allows agents to directly modulate their norm sensitivities as a function of their possible actions. In fact, as modeled in the simulations, *the two models conditionalize on different criteria*. The KV model expresses norm sensitivity conditions on the *past behavior* via ρ_i, but the CGT model expresses norm sensitivity by conditioning on *current conjectures* as functions of explicitly defined social relationships. Although the CGT approach certainly permits conditioning on the history, we have deliberately avoided doing so in order to emphasize conditioning via direct social influence.

4 Discussion

In this section we sketch the wider theoretical framework that the modeling and simulation exercise of the previous two sections is intended to exemplify. We emphasise that this is a sketch: substantial technical work necessary to develop it into theory lies ahead.

We understand Population A in the simulation as a community in which the only strategic interactions occur between agents with fixed utility functions. Population B illustrates a community in which interactions occur under the shadow of a background process of social normative pressure that alters utility functions through mindshaping. Such processes do not necessarily tend toward homogeneity; two pairs of individuals that begin with conflicting preferences from one another might, in developing closer within-pair preference structures, become less aligned between pairs. Mindshaping can thus promote polarisation, or balkanisation in parts of networks, in addition to normative convergence in other parts. Ross, Stirling and Tummolini (2023) simulate such cases, along with cases of preference falsification following Kuran (1995), in a setting where outcomes differ from one another only in relative risk. Which of these general kinds of social situations mindshaping produces depends on the structure of the network, which our work to date demonstrates but for which we have yet to develop general theory. We conjecture that in relatively simple networks such as the one simulated above, in which individuals mainly play games in which Pareto-dominant coordinated equilibria are available, mindshaping will tend to accelerate the emergence of efficiency, as in the simulation.

One potential application of CGT modeling of mindshaping can occur in the experimental lab when subjects encounter novel situations to which pre-established norms do not apply. Experimenters in such situations often run pre-play phases to help agents learn about the game (Fudenberg and Levine 1998). Under a well-explored range of conditions, such learning allows Bayesian agents to identify correlated equilibria (Aumann 1974, 1987). Ross and Stirling (2023) show that mindshaping as modeled by CGT is among these conditions.

Our application of CGT and mindshaping to the KV experiment, however, illustrates a more novel and ambitious modeling vision for relating long-run and short-run strategic processes. The idea envisages two kinds of models, at different scales of representation, that constrain one another through shared use of non-refined non-cooperative game theory. By 'non-refined' we refer to game theory that aims to *describe* patterns of strategic behavior rather than *prescribe* it according to idealisation of stringently rational choice. Of course, game theory cannot be applied to entirely irrational or arational behavior. Agents' choices must be sensitive to changes in incentives that can be identified using utility functions. Over interesting explanatory domains of interaction, models must predict, and risk refutation by failure of, stochastic dominance. Equilibria no weaker in the constraints they impose on identification than quantal response equilibrium (QRS) must be supported by analysis. Taken together, these constraints amount, in philosophical terms, to the idea that most accurate and comprehensive explanation and generalisation of agents' behavior *requires* adoption by the modeler of the intentional stance (Dennett 1987, Ross 2005, 2014).

We aim to *supplement* but not *displace* 'conservative' modeling that identifies agents by means of fixed utility functions, thus enabling extensive-form representations of games and applications of standard non-cooperative solution concepts (e.g. best-reply, Nash, Bayes-Nash, or Quantal Response equilibria). In the conservative setting, agents need not be expected utility maximizers, but their choices should be identifiable using axioms that formally nest expected utility theory (EUT), such as rank-dependent utility (RDU) theory (Quiggin 1982) and dual theory (Yaari 1987). [4]

Among models of the effects of social norms, both the general and special models of Kimbraugh and Vostroknutov (2016 and 2020a, respectively) count as conservative according the above restrictions. This also applies to many models in the social-preferences tradition, notwithstanding our endorsement of Binmore's (2016) complaint that this approach eschews generality and therefore poorly serves the goal of accumulation of knowledge in theory. In holding utility functions fixed, these models are best interpreted as describing strategic interactions of agents for which the distribution in a population of preferences, on which norms supervene, are exogenous. KV, as noted, do not aspire to describe processes by which norms originate and change. In our setting, we interpret this as reflecting the idea that what turns some agents' preferences into norms are diffusion (mindshaping) processes of the kind illustrated in our CGT simulation.

[4] This excludes prospect theory, which may in its original formulation (Kahneman and Tversky 1979) have merits as a model of some psychological processes, but in its economic expression as cumulative prospect theory (Tversky and Kahneman 1992) adds elements we regard as ad hoc (see Harrison and Swarthout 2023).

Of course there are many approaches one could imagine taking to modeling mindshaping other than CGT. Our specific aim is to represent a *strategic* aspect of mindshaping. This is consistent with a range of approaches to social theory that views individuals as strategically responsive to norms as social facts, in ways that dynamically transform these facts but often in ways that no agent intends. We have recurrently cited Kuran (1995) as an example, but more general treatments can be found in Coleman (1990) and Martin (2009); see Ross (2014) for methodological and philosophical discussion. The key feature that differentiates macrostructural models of normative influence from microeconomic models is that the former represent preferences and utility functions as *socially adaptive*. Such processes cannot be captured as solutions to extensive-form games among fixed agents. CGT handles this limitation not by adopting novel solution concepts, but by applying standard solution concepts only after applying marginalisation to games played among individuals with conditional utility functions, according to the procedures described in Section 3.

Adaptive preferences have been presented by various authors following Sen (1992) as challenges to standard welfare theory (see also Elster 1983). Part of the problem is mere intertemporal inconsistency. Where this is not a problem, perhaps because time-scales are short, we see no motivation to supplement the tool-kit we called ‘conservative’. But a more substantive basis for concern is that individual preference adaptation as an empirical phenomenon often involves less powerful agents accommodating the objectives of more powerful ones. In a tradition in which the centrality of welfare as a proxy for well-being is based on emphasizing consumer sovereignty, and that in turn rests on aversion to paternalism by a morally non-neutral state, preference adaptation that results from power imbalance strikes many analysts as sinister.

With respect to these normative issues, we note that our simulation of the public goods game in a CGT setting is a reminder of the morally ambiguous nature of mindshaping. It seems to us to be an obvious historical fact that large-scale normative change has very frequently reflected cultural imperialism and coercive homogenisation. On the other hand, to the extent that such homogenisation expands the moral circle, in the sense of Singer (1981), by promoting the spread of normative sensitivity across sub-networks, it is likely the essential basis for values that aspire to universal scope. In the more modest context of a public goods game, as our simulation demonstrates, it can accelerate and stabilise cooperative choice. We furthermore suggest that by directly representing the *strategic* element of social preference adaptation, CGT offers a technique for correcting some accounts of political and cultural dynamics that neglect the agency of the oppressed.

Our modeling exercise leaves the general, formal integration of the macrosociological and microeconomic dynamics of norms as a pending project. We argue, however, that it displays a promising technical tool for the project, and begins to specify its conditions. We close the discussion with a sketch of the technical agenda we have in mind.

Our template for the proposed relationship between CGT and standard game theory is the relationship between cooperative and non-cooperative game theory as conceived in the Nash program. Binmore (1998) provides a rich informal discussion. Since we are here outlining a proposed methodology rather than yet building real theory, this informal exposition is the appropriate reference point.

Nash (1950, 1951, 1953) identified the value of a practical strategy, for use in applied work, that could exploit the strengths of both cooperative and non-cooperative game theory while avoiding their respective weaknesses. Cooperative game theory can identify highly general solutions to bargaining games that are robust to many changes in details about players’ utility functions and strategy sets. However, none of the various cooperative solution concepts that have been studied have been found to be robust across all bargaining games. By contrast, solution sets for non-cooperative games can be identified with confidence, and it is typically clear from a game’s structure which solution concept is most appropriate for application. One might therefore imagine that when God applies game theory to the strategic affairs of her creatures she doesn’t bother with cooperative models: she writes down the full non-cooperative extensive form of the bargaining game among the individuals who are conjectured into coalitions by the less powerful minds of mortal theorists. She then computes the equilibria of the game that folds both the ‘pre-play’ and ‘play’ stages into a single extensive-form model with stages. Unfortunately, theorists attempting to construct applied models of real-world situations without unbounded cognitive resources typically can have little confidence that they have reliably sorted relevant from irrelevant differences among players and strategy sets when they try to reconstruct the pre-play cooperative game in non-cooperative extensive form. Such sorting is essential to avoid combinatorial explosions of branches in game trees.

Nash's method for coping with this problem - the Nash program - is to use cooperative analysis as the basic tool for representing bargaining games, but then to use non-cooperative analysis to test proposed cooperative solutions. The straightforward idea is that the modeler should be obliged, after identifying a cooperative solution, to find a plausible non-cooperative model for which the solution in question is an equilibrium in extensive form. Nash illustrated the strategy in defending his solution to the general problem of bargaining over the distribution of a cooperatively generated surplus.

We suggest a similar approach to reconciling macrostructural and microeconomic analysis as we have characterized them here. This is the approach we illustrated in Section 3 as applied to KV's public goods game.

Consider first a theorist whose primary interest is in the dynamics of normative stability and change in a society or sub-society. She can begin by building a conditional game to describe the network of relationships among types of agents that she hypothesises on empirical grounds. But actual interactions involve agents with situation-specific parameters in their utility functions. To ensure that she will be able to compare her model with results of this more granular and particular modeling, our social theorist should avoid various restrictions that would limit generality. For example, she should not impose homogenous risk attitudes on her population or assume that all agents are expected-utility maximisers. This she can do by assigning them response functions with flexible (e.g., Prelec) decision weights, following the approach in Ross, Stirling and Tummolini (2023). But there is one key restriction the conservative game theorist makes that she drops: she allows agents' utility functions (and their decision weights on risky prospects) to be influenced through conditionalisation and marginalisation as Markov processes.

Though the empirical process that is the social theorist's explanatory target is dynamical, her CGT model is analysed as comparative statics. Use of CGT allows her to restrict her solution set to concepts that are as narrow as her favoured philosophy of game theory recommends; we would urge that she follow the advice of Binmore (2007) in making this choice. The structural analyst can then test whether the social normative equilibrium she has identified is stable across standard game-theoretic specifications. By this we mean that the equilibrium is reachable in an extensive-form unconditional model of her game, played by agents with utility functions 'frozen' in the underlying social dynamics.

The importance of explicitly attending to norms in a unified social analysis emerges at this methodological juncture. The point of the CGT modeling is to identify them. If the extensive-form unconditional games against which they are tested does not allow for norm-responsiveness in players' utility functions, then there is no reason to expect that the conditional and unconditional solutions should align. Thus the unconditional game used in the test that is analogous to that recommended by the Nash program should be defined using utility functions such as those developed by KV (2020a).

Now let us examine this methodological program from the microeconomist's point of view. Her aim is to predict and explain outcomes of interactions among individual agents who are identified by their utility functions. If these agents are people or human institutions (or even, perhaps, elephants or orcas or ravens), their assessments of outcomes will be (heterogeneously) influenced by prevailing social norms. These norms are exogenous to the microeconomist's model: her agents did not choose them. She might try to account for them by conjecturing and experimenting with various social-preferences models. But this is likely to be extremely inefficient, and in any event is likely to generate over-fitted models even if she finds a social-preferences conjecture that happens to 'work', because her agents' utility functions will be more restricted than those that characterise the 'frozen' output of the corresponding CGT analysis. In the KV setting, what matters is not which precise social states different players idiosyncratically prefer; rather, what she seeks to take into account is the distribution of responsiveness among her agents to whichever norms prevail and must be factored into their empirical expectations about the strategies they will encounter.

In the public goods game example we simulated, under conditions where enough economic agents are sufficiently sensitive to a norm of cooperation to find equilibria, mindshaping processes can accelerate the extent of entrenchment of the norm over time. However, the threshold frequency of norm-sensitivity sufficient for increasing the frequency of cooperative play is substantially higher than the threshold required for a stationary non-zero distribution of cooperative play among agents without mindshaping. This result is intuitively appealing, suggesting a general micro-mechanism for the kinds of cultural ratchet effects that promote the spread of political participation, rule of law, and more globalised markets described by theorists of the historical spread of democratic capitalism, and of contemporary countries' relative levels and rates of success in what Fukuyama (2014) calls 'getting to Denmark' (Lal 1998; Grief 2006; Henrich 2020; Acemoglu and Robinson 2012).

Mindshaping also allows for representation of processes of a kind that standard game-theoretic models do not aim to explain. Preference falsification or pluralistic ignorance about norms are sustainable as equilibria among agents with fixed preferences (Smerdon, Offerman & Gneezy 2019). However, the fixation of preferences blocks endogenisation of the kind of process described by Kuran (1995) in which preference falsification ends not with general discovery of social error, but with adaptation of preferences to comply with observed choices.

As discussed in Section 3, this has somewhat vertiginous implications for welfare analysis. General discovery of preference falsification should tend to increase welfare, by bringing both individual behaviour and policy into closer alignment with preferences. Elimination of preference falsification through preference adaptation should *also* tend to promote welfare improvement, but by changing preferences rather than behaviour or policy. This does not necessarily carry societies in the direction away from Denmark - think of a society with an initial majority of private racists who over time adjust their preferences to conform with public normative shaming - but we see no prima facie basis for general optimism. For example, the Chinese government might well have succeeded over the past couple of decades in creating more sincerely militant nationalism as a means of shoring up the legitimacy of its authoritarian institutions. Theorists who insist on subjective preference satisfaction as the touchstone for welfare assessment might have to acknowledge that the recent Chinese repression of Hong Kong's autonomy may thus have improved Chinese social welfare, as measured against actual preferences in the general population. We do not intend this comment as a kind of passive-aggressive attack on welfarism. Perhaps it is a reminder of the virtues of welfare criteria that are often criticised as being too conservative. No matter how many nationalistic Chinese people have enjoyed better satisfied preferences through the assault on rule of law in Hong Kong, it is clearly no Pareto improvement, though it might satisfy Kaldor-Scitovsky criteria. Recent government behaviour in Xinjiang can constitute neither sort of welfare improvement, as the losers cannot be compensated by any feasible lump-sum transfer.

A final discussion issue involves conceptual sanitation. We earlier suggested that Ross's (2006, 2008) overlapping-generations model of what he calls 'game determination' should be displaced by the more encompassing strategy adopted here. However, the general phenomenon of strategically influenced social-psychological change that Ross (2005) describes informally but more richly remains the explanatory target. We therefore adopt the following semantic policy going forward: mindshaping is the basic mechanism for game determination, of which CGT is a more general formal theory than Ross's model.

5 Conclusion

Constructing a theory of social norms is an ambitious, developing enterprise that will necessarily be the work of many hands from multiple disciplines. We have sketched a methodology that aims to facilitate that interdisciplinarity without sacrificing the accumulation of analytic clarity that the traditional division of labour in the social sciences has fostered. We follow the advice of Gintis (2009) in using game theory for social-scientific unification. But we also follow Binmore (1994, 1998, 2005) in retaining the straight and narrow path of standard microeconomic theory in modeling shorter-run games among socialised individuals. Modeling longer-run dynamics is a task for social theory more generally. Then the crucial trick for the promoter of unification to turn is finding a technical meta-language in which mutual constraints from models at different scales of abstraction can be stated. We proposed conditional game theory as furnishing a possible such meta-language.

Norms are produced by mindshaping processes that are the critical mechanism explaining human ecological dominance (Henrich 2015). These norms are exogenous features of the contexts in which individual economic agents encounter one another as traders and bargainers. We see these as the basis for fundamental axioms of a future theory of norms.

References

Acemoglu, Daron, & Robinson, James. 2012. *Why Nations Fail*. Crown.

Akerlof, George, & Kranton, Rachel. 2010. *Identity Economics*. Princeton University Press.

Andreoni, James, & Bernheim, B.Douglas. 2009. Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica* 77: 1607-1636.

Andreoni, James, Nikiforakis, Andreoni, & Siegenthaler, Simon. 2017. Social change and the conformity trap. WP, Semantic Scholar: https://api.semanticscholar.org/CorpusID:53371491

Aumann, Robert 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1: 67-96.

Aumann, Robert 1987. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica* 55: 1-18.

Bernheim, B.Douglas. 1994. A theory of conformity. *Journal of Political Economy* 102: 841-877.

Bicchieri, Cristina. 2006. *The Grammar of Society*. Cambridge University Press.

Bicchieri, Cristina. 2017. *Norms in the Wild*. Oxford University Press.

Binmore, Ken. 1994. *Game Theory and the Social Contract, Volume 1: Playing Fair*. MIT Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract, Volume 2: Just Playing*. MIT Press.

Binmore, Ken. 2005. *Natural Justice*. Oxford University Press.

Binmore, Ken. 2007. *Playing For Real*. Oxford University Press.

Binmore, Ken. 2010. Social norms or social preferences? *Mind and Society*, 2: 139-157.

Bisin, Alberto, & Verdier, Thierry. 2001. The economics of cultural transmission and the dynamics of preferences. *Journal of Economic Theory* 97: 298-319.

Bogdan, Radu. 1997. *Interpreting Minds*. MIT Press.

Boyd, Robert, & Richerson, Peter. 1985 *Culture and the Evolutionary Process*. University of Chicago Press.

Brock, William, & Durlauf, Steven. 2001. Discrete choices with social interactions. *Review of Economic Studies* 68: 235-260.

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

Clark, Andy. 1997. *Being There*. MIT Press.

Coleman, James. 1990. *Foundations of Social Theory*. Cambridge, MA: Harvard University Press.

Davis, John. 2010. *Individuals and Identity in Economics*. Cambridge University Press.

Dennett, Daniel. 1987. *The Intentional Stance*. MIT Press.

Edgerton, Robert. 1992. *Sick Societies*. Free Press.

Elster, Jon. 1983. *Sour Grapes*. Cambridge University Press.

Fehr, Ernst, & Gächter, Simon. 2000. Cooperation and punishment in public goods games. *American Economic Review* 90: 980-994.

Fehr, Ernst & Schmidt, Klaus M. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* 114(3): 817-868.

Fudenberg, Drew, & Levine, David 1998. *Theory of Learning in Games*. MIT Press.

Fukuyama, Francis. 2014. *Political Order and Political Decay: From the Industrial Revolution to the Present Day*. Farrar, Straus and Giroux.

Gilbert, Margaret. 1989. *On Social Facts*. Princeton University Press.

Gintis, Herbert. 2009. *The Bounds of Reason*. Princeton University Press.

Gintis, Herbert. 2016. *Individuality and Entanglement*. Princeton University Press.

Goeree, Jacob, Holt, Charles, & Palfrey, Thomas. 2016. *Quantal Response Equilibrium*. Princeton University Press.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. Anchor.

Grief, Avner. 2006. *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade*. Cambridge University Press.

Harrison, Glenn W., & Swarthout, J.Todd. 2023. Cumulative prospect theory in the laboratory: A reconsideration. In G.W. Harrison and D. Ross, eds., *Models of Risk Preferences: Descriptive and Normative Challenges*. Emerald, pp. 107-192. .

Harsanyi, John. 1967. Games With incomplete information played by 'Bayesian' players, Parts I-III. *Management Science* 14: 159–182.

Henrich, Joseph. 2015. *The Secret of Our Success*. Princeton University Press.

Henrich, Joseph. 2020. *The WEIRDEST People in the World*. Farrar, Straus and Giroux.

Kahneman, Daniel, & Tversky, Amos. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263-292.

Kessler, Judd, & Leider, Stephen. 2012. Norms and contracting. *Management Science* 58: 62-77.

Kimbraugh, Erik, & Vostroknutov, Alexander. 2013. Norms make preferences social. Working paper CiteSeerX: https://www.parisschoolofeconomics.eu/IMG/pdf/vostroknutov_paper.pdf

Kimbrough, Erik, & Vostroknutov, Alexander. 2016. Norms make preferences social. *Journal of the European Economic Association* 14: 608-638.

Kimbrough, Erik, & Vostroknutov, Alexander. 2020a. A theory of injunctive norms. Working paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3566589

Kimbrough, Erik, & Vostroknutov, Alexander. 2020b. Injunctive norms and moral rules. Working paper. http://www.vostroknutov.com/pdfs/axinorms3_00.pdf

Kreps, David, & Wilson, Robert. 1982. Sequential equilibria. *Econometrica* 50:863-894.

Krupka, Erin, & Weber, Roberto. 2013. Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association* 11: 495-524.

Kuran, Timur. 1995. *Private Truths, Public Lies*. Harvard University Press.

Lal, Deepak. 1998. *Unintended Consequences*. MIT Press.

Ledyard John. 1995. Public goods: a survey of experimental research. In H. Kagel & J. Roth, eds., *Handbook of Experimental Economics*, pp. 253– 279. Princeton University Press.

Martin, John. 2009. *Social Structures*. Princeton: Princeton University Press.

McClamrock, Ron. 1995. *Existential Cognition*. University of Chicago Press.

Michaeli, Moti, & Spiro, Daniel. 2015. Norm conformity across societies. *Journal of Public Economics* 132: 51-65.

Michaeli, Moti, & Spiro, Daniel. 2017. From peer pressure to biased norms. *American Economic Journal: Microeconomics* 9: 152-216.

Nash, John. 1950. The bargaining problem. *Econometrica* 18: 155-162.

Nash, John. 1951. Non-cooperative games. *Annals of Mathematics* 54: 286-295.

Nash, John. 1953. Two-person cooperative games. *Econometrica* 21: 128-140.

Nichols, Shaun, & Stich, Stephen. 2003. *Mindreading*. Oxford University Press.

Oxoby, Robert, & Spraggon, John. 2008. Mine and yours: Property rights in dictator games. *Journal of Economic Behavior and Organization* 65: 703-713.

Quiggin, John. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3: 323–343.

Quiggin, John. 1993. *Generalized Expected Utility Theory. The Rank-Dependent Model*. Boston: Kluwer.

Ross, Don. 2005. *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.

Ross, Don. 2006. The economics and evolution of selves. *Cognitive Systems Research* 7: 246-258.

Ross, Don. 2008. Classical game theory, socialization, and the rationalization of convention. *Topoi* 27: 57-72.

Ross, Don. 2014. *Philosophy of Economics*. Palgrave Macmillan.

Ross, Don, & Stirling, Wynn. 2021. Economics, social neuroscience, and mindshaping. In J. Harbecke & C. Herrmann-Pillath, eds., *Social Neuroeconomics*, pp. 174-201. Routledge.

Ross, Don, & Stirling, Wynn. 2023. Mindshaping, conditional games, and the Harsanyi doctrine. CEAR Working Paper 2023-03.

Ross, Don, Stirling, Wynn, & Tummolini, Luca. 2023. Strategic theory of norms for empirical applications in political science and political economy. In H. Kincaid & J. Van Bouwel, eds., *The Oxford Handbook of Philosophy of Empirical Political Science*. Oxford University Press, pp. 86-121.

Samuelson, Paul. 1958. An exact consumption-loan model of interest with or without the social contrivance of money. *Journal of Political Economy* 66: 467–482.

Sen, Amartya. 1992. *Inequality Reexamined*. Harvard University Press.

Singer, Peter. 1981. *The Expanding Circle*. Farrar Straus & Giroux.

Smerdon, David, Offerman, Theo, & Gneezy, Uri. 2019. 'Everybody's doing it': On the persistence of bad social norms. *Experimental Economics* 23: 392-420.

Smith, Vernon, & Wilson, Bart. 2019. *Humanomics*. Cambridge University Press.

Stirling, Wynn. 2012. *Theory of Conditional Games*. Cambridge University Press.

Stirling, Wynn. 2016. *Theory of Social Choice on Networks*. Cambridge University Press.

Sugden, Robert. 1998. Normative expectations: the simultaneous evolution of institutions and norms. In A. Ben-Ner , & L. Putterman, eds., *Economics, Value, and Organization*, pp. 73–100. Cambridge University Press.

Sugden, Robert. 1986/2004 *The economics of rights, co-operation and welfare*. Palgrave Macmillan.

Tummolini, Luca, & Stirling, Wynn. 2020. Coordinated rational choice. *Topoi* 39(2): 317-327.

Tversky, Amos, & Kahneman, Daniel. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5: 297–323.

Weibull, Jorgen. 1995. *Evolutionary Game Theory*. MIT Press.

Yaari, Menahem. 1987. The dual theory of choice under risk. *Econometrica* 55: 95-115.

Yamagishi, T. 1986. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51: 110-116.

Young, Peyton H. 2015. The evolution of social norms. *Annual Review of Economics* 7: 359-387.

Zawidzki, Tadeusz. 2013. *Mindshaping*. MIT Press.

Appendix

A Conditional Game Theory Review

A.1 Definitions and Notation

Definition A.1. *An* influence network graph $G(\mathbf{X}, E)$ *comprises a set of* vertices $\mathbf{X} = \{X_1, \ldots, X_n\}$ *(the set of agents) and a set* $E \subset \mathbf{X} \times \mathbf{X}$ *of pairs of vertices such that there is an explicit connection between them that serves as the medium by which influence is propagated between* X_i *and* X_j*. The expression* $X_i \longrightarrow X_j$ *means that the influence propagates in only one direction—a* directed edge *from* X_i *to* X_j*. A* path *from* X_j *to* X_i *is a sequence of directed edges from* X_j *to* X_i*, denoted* $X_j \mapsto X_i$*. A*

For each X_i*, its* parent *set is* $\mathrm{pa}\,(X_i) = \{X_{i_1}, \ldots X_{i_{q_i}}\}$*, where* $X_{i_k} \longrightarrow X_i$*,* $k = 1, \ldots, q_i$*. A graph is said to be* directed *if all edges are directed; it is a* directed acyclic graph *if all edges are directed and there are no cycles. If* $\mathrm{pa}\,(X_i) = \varnothing$ *then* X_i *is a* root vertex. *A directed graph is a* cyclic directed graph *if there are no root vertices.*

Definition A.2. *A* conditional network game *is a triple* $\{\mathbf{X}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{U}}\}$*, where* $\mathbf{X}$ *is the set of agents;* $\mathcal{A}_i = \{x_{i1}, \ldots, x_{iN_i}\}$*,* $i = 1, \ldots, n$*, is the set of* actions *available to* X_i*;* $\boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ *is the set of* outcomes*; and* $\boldsymbol{\mathcal{U}} := \{u_{i|\mathrm{pa}(i)},\ i = 1, \ldots, n\}$ *is the set of* conditional utilities *such that* $u_{i|\mathrm{pa}(i)}$ *is the utility to* X_i *as modulated by its conjectures regarding the actions taken by its parents.*

Definition A.3. *A* self-conjecture *for* X_i*, denoted* $X_i \vDash a_i$ *for* $a_i \in \mathcal{A}_i$*, is an action under consideration by* X_i *for implementation. For* $X_{i_k} \in \mathrm{pa}\,(X_i)$*, a* conditioning conjecture *by* X_i *for* X_{i_k}*, denoted* $X_{i_k} \vDash a_{i_k}$ *for* $a_{i_k} \in \mathcal{A}_k$*, is an action that* X_i *hypothesizes that* X_{i_k} *is considering for implementation,* $k = 1, \ldots, q_k$*. A* conditioning conjecture set $\boldsymbol{\alpha}_{\mathrm{pa}(i)} = (a_{i_1}, \ldots, a_{i_{q_i}})$ *for* $\mathrm{pa}\,(X_i)$ *is the set of conditioning conjectures by* X_i *for its parents, denoted* $\mathrm{pa}\,(X_i) \vDash \boldsymbol{\alpha}_{\mathrm{pa}(i)}$.

Definition A.4. *A* conjecture hypothesis*, denoted*

$$\mathcal{H}_{i|\mathrm{pa}(i)}(a_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)})\text{:}\ \mathrm{pa}\,(X_i) \vDash \boldsymbol{\alpha}_{\mathrm{pa}(i)} \Longrightarrow X_i \vDash a_i \tag{A.1}$$

is a hypothetical proposition that, if $\boldsymbol{\alpha}_{\mathrm{pa}(i)}$ *is a conditioning conjecture set for* $\mathrm{pa}\,(X_i)$ *(the antecedent), then* X_i *will conjecture* a_i *(the consequent). A* conditional utility *given* $\boldsymbol{\alpha}_{\mathrm{pa}(i)}$*, denoted* $u_{i|\mathrm{pa}(i)}(\cdot|\boldsymbol{\alpha}_{\mathrm{pa}(i)})$*, is an ordering function such that, given the antecedent* $\mathrm{pa}\,(X_i) \vDash \boldsymbol{\alpha}_{\mathrm{pa}(i)}$*, then*

$$u_{i|\mathrm{pa}(i)}(a_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)}) \geqslant u_{i|\mathrm{pa}(i)}(a_i'|\boldsymbol{\alpha}_{\mathrm{pa}(i)}) \tag{A.2}$$

means that the consequent $X_i \vDash \mathbf{a}_i$ *is either strictly preferred to the consequent* $X_i \vDash a_i'$ *or* X_i *is indifferent, given that its parents conjecture* $\boldsymbol{\alpha}_{\mathrm{pa}(i)}$*. If* $\mathrm{pa}\,(X_i) = \varnothing$*, then* $u_{i|\mathrm{pa}(i)}(a_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)}) = u_i(a_i)$*, a* categorical utility.

Since utilities are invariant with respect to positive affine transformations, it may be assumed without loss of generality that the conditional utilities are nonnegative and sum to unity; that is,

$$\begin{aligned} u_{i|\mathrm{pa}(i)}(a_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)}) &\geqslant 0 \text{ for all } a_i \in \mathcal{A}_i \\ \sum_{a_i} u_{i|\mathrm{pa}(i)}(a_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)}) &= 1 \text{ for all} \boldsymbol{\alpha}_{\mathrm{pa}(i)}\,. \end{aligned} \tag{A.3}$$

These definitions correspond to a special case of conditional game theory as originally introduced in Stirling (2012). With general conditional game theory, the conditional utilities are mappings $u_{i|\mathrm{pa}(i)}\text{:}\boldsymbol{\mathcal{A}}|\boldsymbol{\mathcal{A}}^{q_i} \rightarrow [0, 1]$, that is, X_i defines its utility over the outcome set (as does standard game theory) conditioned on outcome conjectures for all of its parents. This formulation is a generalization of noncooperative game theory, and degenerates to a standard noncooperative game if no agent conditions on other agents—a network with no edges. However, since our study involves only the special case, we confine our discussion accordingly.

A.2 Acyclic Conditional Game Model

Conditional game theory applies syntactical structure of Bayesian network theory with agents (analogous to random variables) as vertices and edges as conditional utility functions (analogous to conditional probability mass functions) that convey social influence from the parents to the children. Analogous to the way the conditional mass functions are combined via the chain rule to generate a joint probability mass functions, the conditional utilities are combined via the chain rule to generate a *coordination function* that captures all of the nascent social relationships that emerge as the agents interact (cf. Pearl (1988), Stirling (2012, 2016)). Thus, the coordination function comprises the product of the conditional utility mass functions, yielding

$$w_{1:n}(a_1, \ldots, a_n) = \prod_{i=1}^{n} u_{i|\mathrm{pa}(i)}(a_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)})\,, \tag{A.4}$$

where $(a_1, \ldots, a_n)$, termed the *coordination profile*, is the set of self-conjectures of $\{X_1, \ldots, X_n\}$. If $u_{i|\mathrm{pa}(i)}(a_i) = u_i(a_i)$, a categorical utility, if $\mathrm{pa}\,(X_i) = \varnothing$ (i.e., X_i is a root vertex).

The individual coordinated utility functions are obtained by marginalization, yielding

$$w_i(a_i) = \sum_{\neg a_i} w_{1:n}(a_1, \ldots, a_n)\,, \tag{A.5}$$

where the notation $\sum_{\neg a_i}$ defines the *exclusion sum*–the sum is taken over all elements in the argument list *except* a_i.

CGT thus appropriates all of the syntactical machinery of probability theory, but with different semantics. Analogous to the way a joint probability mass function serves as a comprehensive model of the statistical interrelationships among a collective of random variables, the coordination function serves as a comprehensive model of the social interrelationships among a collective of agents. It provides a ranking of the degrees of compatibility for all action profiles and characterizes the propensity of the members of the network to behave in a systematic and organized way. Whereas the conditional utility $u_{i|\mathrm{pa}(i)}$ provides an *ex ante* conditional ordering over X_i's action set before social interaction occurs, the coordinated utility w_i provides an *ex post* ordering after having taken into consideration the effects of social interaction.

A.3 Extension to Cyclic Networks

The conditional game model may be extended to include cyclic influence of the form

$$\begin{tikzcd} X_1 \arrow[r, bend left, "u_{2|1}"] & X_2 \arrow[l, bend left, "u_{1|2}"] \end{tikzcd} \tag{A.6}$$

by viewing this scenario as an infinite sequence of interrelationships that occur as time evolves, where X_1 influences X_2 who then influences X_1, who again influences X_2, and so forth. The central issue is whether such a sequence of transitions oscillates unendingly or ultimately converges to a steady state of fixed utilities for each agent. Fortunately, however, since CGT complies with the syntax of probability theory, we may apply Markov chain convergence theory to address this scenario.

In a standard probability context, a discrete-time Markov process is a sequence of time-indexed random variables $\{Y(\tau), \tau \in \{1, 2, \ldots\}$ of the form

$$\begin{tikzcd} Y(1) \arrow[r, "p_{2|1,\,\tau=1}"'] & Y(2) \arrow[r, "p_{3|2,\,\tau=2}"'] & Y(3) \arrow[r, "p_{4|3,\,\tau=3}"'] & Y(4) \arrow[r, "p_{5|4\,,\tau=4}"'] & \cdots\,, \end{tikzcd} \tag{A.7}$$

where $p_{\tau+1|\tau}$ is the conditional probability mass function governing $Y(\tau+1)$ given $Y(\tau)$. This probability structure assures that $Y(\tau-1)$ and$Y(\tau+1)$ are conditionally independent, given $Y(\tau)$. In other words, the Markov property is equivalent to the statement that the state of the past and the state of the future are conditionally independent, given the state of the present.

Analogously, we may view the network defined by (A.6) as a collective of time-sequenced acyclic networks of the form

$$X_1(1) \xrightarrow[u_{2|1,\,\tau=1}]{} X_2(2) \xrightarrow[u_{1|2,\,\tau=2}]{} X_1(3) \xrightarrow[u_{2|1,\,\tau=3}]{} X_2(4) \xrightarrow[u_{1|2\,,\tau=4}]{} \cdots\,. \tag{A.8}$$

Definition A.5. *The agents $X_1(\tau - 1)$ and $X_1(\tau + 1)$ are* conditionally socially independent*, given $X_2(\tau)$, if the the conditional subgroup coordination function satisfies the condition*

$$w_{\tau-1,\tau+2|\tau}(a_1, a_1'|a_2) = w_{\tau-1|\tau}(a_1|a_2)w_{\tau+1|\tau}(a_1'|a_2)\,. \tag{A.9}$$

We express this condition with the notation $X_1(\tau - 1)\perp X_1(\tau + 1)|X_2(\tau)$.

Suppose at iteration $\tau = 1$, X_1's marginal utility is $w_1(a_1, 1)$ (with the second argument corresponding to iteration), the coordination function at iteration $\tau = 2$ is, applying (A.4),

$$w_{12}(a_1, a_2, 2) = w_1(a_1, 1)u_{2|1}(a_2|a_1)\,, \tag{A.10}$$

with marginal for X_2 computed at iteration $\tau = 2$ using, as (A.5) as

$$w_2(a_2, 2) = \sum_{a_1} w_{12}(a_1, a_2, 2)\,. \tag{A.11}$$

The coordination function and marginalization may be combined using matrix notation

$$\mathbf{w}_i(\tau) = T_{i|j}\mathbf{w}_j(\tau)\,, \tag{A.12}$$

where the *mass vector* is

$$\mathbf{w}_i(\tau) = \begin{bmatrix} w_i(x_{i1}, \tau) \\ w_i(x_{i2}, \tau) \\ \vdots \\ w_i(x_{iN_i,\tau}) \end{bmatrix} \tag{A.13}$$

and

$$T_{i|j} = \begin{bmatrix} u_{i|j}(x_{i1}|x_{j1}) & \cdots & u_{i|j}(x_{i1}|x_{jN_j}) \\ \vdots & & \vdots \\ u_{i|i}(x_{iN_i}|x_{i1}) & \cdots & u_{i|j}(x_{iN_i}|x_{jN_j}) \end{bmatrix} \tag{A.14}$$

is the *state-to-state transition matrix* from X_j to X_i for $i|j \in \{1|2,\ 2|1\}$. Thus, we may express the state of X_i at iteration τ as

$$\mathbf{w}_i(\tau) = T_{i|j}\mathbf{w}_j(\tau - 1) = T_{i|j}T_{j|i}\mathbf{w}_i(\tau - 2) = T_i\mathbf{w}_i(\tau - 2)\,, \tag{A.15}$$

where $T_i = T_{i|j}T_{j|i}$ is the *closed-loop transition matrix*. In general, it holds that

$$\mathbf{w}_i(\tau) = T_i^{\tau}\mathbf{w}_i(0)\,, \tag{A.16}$$

where τ is now expressed in closed-loop iteration increments. The key result of Markov theory is the *Markov chain convergence theorem*.

Theorem A.1. *If T is a transition matrix with all entries strictly greater than zero, there exists a unique mass vector $\overline{\mathbf{w}}$ such that a) $T\overline{\mathbf{w}} = \overline{\mathbf{w}}$; b) for any initial state $\mathbf{w}(0)$, the* steady-state mass vector *is*

$$\overline{\mathbf{w}} = \lim_{\tau\to\infty} T^{\tau}\mathbf{w}(0), \tag{A.17}$$

and c)

$$\lim_{\tau\to\infty} T^{\tau} = \overline{T}\,, \tag{A.18}$$

where $\overline{T} = \begin{bmatrix}\overline{\mathbf{w}} & \cdots & \overline{\mathbf{w}}\end{bmatrix}$.

For a proof of this theorem, see Luenberger (1979) or Stirling (2016).

References

Luenberger, David. G. 1979. *Introduction to Dynamic Systems*. John Wiley.
Pearl, Judea. 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman.
Stirling, Wynn. 2012. *Theory of Conditional Games*. Cambridge University Press.
Stirling, Wynn. 2016. *Theory of Social Choice on Networks*. Cambridge University Press.