Strategic Theory of Norms for Empirical Applications in Political Science and Political Economy

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

1 Introduction

The study of social norms sprawls across all of the social sciences. Consequently, the concept lacks a unified conception, let alone a generally acknowledged formal theory. In this chapter we do not seek, grandiosely, to legislate across the various approaches that have been developed. However, we do aim to synthesize an account that can be applied generally, at the *social* scale of analysis, and can be applied to empirical evidence generated in the experimental laboratory and in field experiments.

More specifically, we provide new analysis on representing norms for application in political science, and in parts of economics that do not follow the recent trend among some behavioral economists to build models of the cognitive and motivational states of individuals taken, as it were, one at a time.[1] Sociologists, criminologists, and social psychologists may also find our theoretical construction useful. For purposes of our analysis, a norm is a feature of a social structure (Martin 2009), that is, an element of prevailing patterns in relationships among recurrently interacting people that constrains and motivates the behaviour of at least some of them, which arises, persists for a finite time, and ends by decay or catastrophic collapse. Though a norm depends for its continuing existence on being behaviorally accommodated by a significant subset of agents embedded in a social structure, any given agent coexisting with a norm in her social environment may or may not cognitively recognize the norm in question, and may adapt her behavior to accommodate the norm to a varying degree over time, where the variance in question can range from unwavering commitment to complete neglect.

Prior to embarking on analysis, it may be helpful to indicate examples of norms in the sense we intend. Alongside legal injunctions and other explicit requirements, people in all large-scale human societies are also regulated by more informal social norms against, for instance, self-serving factual misrepresentation, self-interested and institutionally unauthorized coercion of people, and wilfully selfish driving that impedes the efficiency of traffic flow. Norms of most interest to social scientists are often those on which there is variation among people who co-habit geographically and politically. Thus in contemporary Western societies majorities behaviorally and cognitively support, but to varying degrees, a norm of non-discrimination against LGBT people, but there are at the same time very substantial sub-communities in which such dis-

[1] For extended discussion of the intended distinction within economics, see Ross (2014).

crimination is normatively expected. Many norms are usually moralized, such as norms against gratuitous cruelty to animals, but many are not, such as localized norms around appropriate colors for painting houses.

Social scientists are motivated to study norms, and to incorporate the concept of a norm into theoretical models, because the existence and relative strength of norms influences individual and coordinated behavior, affects the sharing and concealment of information, and drives the relative stability of formal and informal institutions. For political scientists, norms are arguably the most important determinants of the efficiency and durability of political orders, as elements of the causal vectors of quintessential political acts like voting (Gerber & Rogers 2009), and contribute to explaining the varieties of democracies and their different systems of deliberation, representation and participation (Mansbridge et al 2006).

Though norms as we understand them here are features of social structures, they have ontological prerequisites at psychological and biological scales. Norms govern the cognition and behavior only of agents - e.g., people, probably some other intelligent social animals, corporations, political parties and lobbies, but not rocks or everyday electronic computers or corpses. Though our account will incorporate some principles of game theory that presuppose various explicit forms of consistency over time in agents, in general our framework assumes only minimal necessary criteria for agency: agents must manifest goal-governed behavior that can be modified by shifts in incentives.

The most important foundational sources for our project in the chapter are Bicchieri (2006, 2017), Kuran (1995), and Stirling (2012, 2016).

From Bicchieri we adapt a general philosophical conception of a norm, according to which a norm exists in a social structure when a significant networked subset of individuals (1) explicitly or implicitly (behaviorally) represent the norm as effective within the subset; and (2) prefer to conform their own behavior faithfully or partially to the norm in instances of social interaction where (2a) they expect that others in the subset (the extension of which may be uncertain) will govern their behavior in accordance with the norm, and (2b) believe that others think that such behavior is what members of the subset *should* do. Bicchieri does not interpret the 'should' here as moral, on grounds that moralized preferences are distinguished by applying unconditionally, and therefore as not depending on one's expectations about others. This implies a philosophically strong and somewhat tendentious conception of morality that we will relax.

From Kuran we take the insight, developed theoretically and based on empirical evidence, that prevailing norms may come to be widely disliked by participants in networks that the norms govern, but nevertheless survive for a time because behavior that would generate public knowledge of this general disenchantment is itself suppressed by the operation of the norm. Such norms are relatively fragile because the suppression of information is likely to leak. This is of special interest in political science, because norms with this characteristic are potential sources of political and social change that appears as sudden and surprising to both participants and observers. At the same time, Kuran's account, unlike most otherwise similar models developed by economists, recognises that publicly expressed preferences for initially widely disfavored norms can feed back upon and modify private preferences. Consequently, his account captures patterns whereby fragile norms that survive into a second generation can manufacture their own climate and become locked in. Incorporation of these insights of Kuran's, as we will formalize them, encourage adjustment in Bicchieri's view of moralised preferences as standing outside of the dynamics of conditionalized norms.

From Stirling we apply *conditional game theory* (CGT). This is an extension of standard noncooperative game theory that incorporates strategic resolution of uncertainty on the part of agents about their own preferences, on the basis of conjecturing and observing evidence about the preferences of others. The explicit motivation for Stirling's theory is to allow game theorists to model the diffusion of social influence through networks as strategically endogenous rather than exogenous. We refer to CGT as an *extension* rather than a *refinement* of standard game theory because it does not narrow the set of standard noncooperative solution concepts by reference to any special model of rationality. The core elements of CGT are summarized in a technical appendix to the chapter.

The chapter is organized as follows. In Section 2 we locate Bicchieri's and Kuran's conceptions of norms

in the wider landscape of concepts used in the social and behavioral sciences, particularly in economics. A main practical purpose of this critical review is to identify tension between Bicchieri's philosophical analysis of norms, which we broadly follow, and the explicit utility functions by means of which she operationalizes that analysis for empirical application. We present Kuran's model of preference falsification dynamics as a natural complement to Bicchieri's analysis of norms that is particularly relevant to the interests of political scientists, and motivate our subsequent replacement of Bicchieri's utility model by Kuran's. In Section 3 we criticise experimental design procedures used by Bicchieri and her co-authors in their applications of her account of norms in the lab, and indicate an improved approach. The choice data to be elicited from such improved procedures imply demands and restrictions on the form of theory required for model identification. The example we use is a multi-player Investment / Trust Game. In section 4 we show how to represent the endogenous resolution of preference uncertainties in an Investment / Trust Game using CGT. In Section 5 we present simulations of two phenomena involving diffusion of normative influence in social networks discussed by both Bicchieri and Kuran. These illustrate the capacity of CGT to represent mechanisms of endogenous norm change, and serve as stylized examples of our procedure for rendering the philosophical account of norms derived from Bicchieri and Kuran as an operational instrument for modeling empirical choice data and for identifying parameters in our enriched version of Kuran's utility model.

Thus the chapter yields the following as sources of value for empirical political scientists: a general philosophical conception of norms as elements of social structure; a high-level experimental method for eliciting attitudes to norms in the lab; a formal theory of norms to aid in writing down empirically identifiable models; and incorporation into the theory of a property of norms, relative fragility, that is fundamental to explaining and perhaps predicting political change.

2 Modelling social norms: categorical versus conditional preferences for conformity

In this section we critically set the theoretical perspectives on norms that we aim to refine and generalize, those of Bicchieri (2006, 2017) and Kuran (1995), in the context of wider literatures.

The construct of social norms figures in most the social sciences – from psychology to sociology, economics and political science – but the lack of a unified, formal, and operational conception has so far limited its use for causal identification and explanation of empirical data. We endorse a widespread view that given game theory's well-developed resources for representing interactional statics and dynamics (Gintis 2014), it provides the most promising technical apparatus for filling this gap.

Since social norms exist insofar as they are complied with, early game-theoretic analyses of norms focused on explaining their characteristic stability. Starting with the seminal contribution of Schelling (1980), norms have been viewed as rules emerging from repeated or recurrent strategic interaction that are stable because they are *self-enforcing*. On this understanding, a necessary condition for something's being a norm is that it must be one among two or more equilibria of a game that is an empirically plausible model of the situation the norm purportedly regulates. More specifically, norms have been characterized as playing the role of equilibrium selection devices (Lewis 1969, Sugden 1986/2004, Binmore 1994, 1998, 2005) or of correlated devices for a correlating equilibrium (Aumann 1987, Gintis 2014, Guala 2016). Although related, the concepts of a social norm and of an equilibrium are not generally coextensive, since two agents might each play their part in an equilibrium that is idiosyncratic to them but not recognized by many or most other participants in the relevant interactive scenario. The stability of a norm critically relies on a shared system of mutual *empirical* expectations, first-order beliefs about the typical behavior of others.

Although her approach is rooted in this tradition, Bicchieri (2006) has argued that the equilibrium conception of social norms naturally fits conventions and other kinds of descriptive norms for which self-interest is enough to motivate conformity (e.g. driving on the left in Cape Town because that is what one expects oth-

ers to do), but falls short of identifying what is peculiar to social norms proper: they often prescribe actions that go against the interests that agents would have independently of the social existence of the norm (e.g., in public I would scratch any part of my body that was itchy if such behavior weren't generally regarded as offensive). According to Bicchieri, social norms arise to regulate behavior in situations characterized both by an element of conflicting interests as well as some potential for general benefit.

Consider for instance the well explored 'Investment' or 'Trust' game first introduced by Berg, Dickhaut and McCabe (1995). In the standard paradigm, an agent, 'the Investor', decides what proportion of an endowment (if any) she will transfer to another agent,'the Trustee'. This action is viewed as investing an amount of money in a project that will generate a surplus, typically simulated in the lab by the experimenter tripling its value. The Trustee decides what proportion of this account to transfer back to the Investor. If players in this game are narrowly self-interested, not risk-lovers, derive utility only over money, and all of this is common knowledge, then no money is transferred in the unique Nash equilibrium of the one-shot game. But players can achieve Pareto superior outcomes if a suitable social norm exists in their society. A norm, for instance, could prescribe that actions should contribute to equality of final monetary positions (an "Equality" norm), or to outcomes that reflect reciprocal proportionality of contributions (an 'Equity" norm), or some other culturally specific norm conditioned on distinctive social roles such as 'parent / child' or 'venture capitalist / entrepreneur'.

The key feature of Bicchieri's analysis is that conformity to whatever norm is established (whether it is Equality or Equity, in the case above) is not primarily driven by one's own intrinsic and stable disposition toward it (i.e. one's personal normative beliefs about what one ought to do) but is instead influenced by how other people in one's reference network are expected to behave and what they believe about one another. In this view, conformity to a social norm is motivated by *conditional preferences*, i.e. an agent prefers to conform to a norm conditional on the fact that (1) she expects that most others will conform to it (*empirical* expectations or first-order beliefs), and (2) that she expects that others believe she ought to conform as well (*normative* expectations or second-order beliefs about the personal normative beliefs of others, that is, what others believe one *ought* to do). It is often added that she must expect sanctioning if she violates the norm, but arguably this is already built into the analysis if one takes a revealed-preference view of expectations, that is, that they must be behaviorally and publicly manifest. Provided that these conditions are met, social norms operate by *transforming* a mixed-motive game like the trust game into a new game in which the interests of norm followers are aligned: norm followers will end up playing a coordination game among themselves where general norm compliance is an equilibrium.

Given the crucial explanatory role conditional preferences play in Bicchieri's framework, an adequate formalization of her analysis should incorporate this concept. First we should distinguish it from two other senses of conditional preference that have featured in the economics literature.

In the first sense, an agent might be said to have a conditional 'preference' for conformity in the norms-as-equilibria conception, because, after all, her reason to conform depends on her expectation about others doing their part in the equilibrium. Expecting different equilibrium behavior from others (and thus a different norm) would motivate a different choice. This 'conditional' preference for a norm-compliant action actually springs from standard fixed and stable preferences defined over outcomes. As clarified by Lewis (1976), an agent can be said to prefer to conform to some rule rather than not, on condition that others conform as well, simply because the state of affairs in which both she and others conform to the rule is preferred to the state of affairs in which others conform but she does not.

A second variety of conditionality is that a preference to conform may depend on the value of an exogenous 'state of nature'. For example, I might prefer to join everyone at the most popular jazz club in town under normal conditions, but find a less well-frequented one if there is an epidemic going around. In this case my preference is state-dependent (Karni 1990; Hirschleifer & Riley 1992; Chambers & Quiggin 2000).

According to Bicchieri, however, neither of these ideas expresses the sense of conditionality that underwrites conformity to social norms. In her view, given the right conditions (the existence of appropriate

empirical and normative expectations), social norms in fact *transform* raw or baseline preferences into new ones. Thus, Bicchieri's conditional preferences are best described counterfactually (Lewis 1976, p. 117), that is, by some kind of hypothetical knowledge of what an agent would prefer if her expectations about others' behavior and beliefs were different. In what follows we do not *assume* that this revisionary understanding of conditional preferences is generally preferable to modeling conditionality as state-dependence. Instead we aim to demonstrate that the revisionary conception is tractable in the formal language of game theory, and under some circumstances adds value as a tool for empirical analysis. Such value potentially includes avoiding the need to resort to an a pre-established ontology determining what does and does not count as a genuinely exogenous 'state of nature' (Andersen et al 2008), a philosophically complex kind of assumption that is implicit in models of state-dependent preference.

In order to formally capture how individual preferences are shaped by the existence of social norms, Bicchieri (2006, 52) initially proposed a model in which the utility function is a linear combination of a player's baseline material payoff and a norm-based component representing the maximum loss suffered by any norm-following player as result of a norm violation. Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ denote a set of n players, $\mathcal{A}_i = \{x_{i1}, \ldots, x_{iM_i}\}$ $i = 1, \ldots, n$ denote their action set and $\mathbf{a} = (a_1, \ldots, a_n) \in \boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ the set of action profiles or outcomes. A norm for a player X_i is represented by a correspondence $\mathcal{N}_i$ from an agent's expectations about the other players' strategies to the strategy the agent *ought* to take, that is $\mathcal{N}_i$: $\mathcal{L}_{-i} \rightarrow \mathcal{A}_i$ with $\mathcal{L}_{-i} \subseteq \mathcal{A}_{-i}$. where $\mathcal{A}_{-i}$ is $\boldsymbol{\mathcal{A}}$ with $\mathcal{A}_i$ removed and a_{-i} is the set $(a_1, \ldots, a_n)$ with a_i removed. A strategy profile $\mathbf{a}$ is said to *violate* a norm when X_j does not follow the norm, that is, when $a_j \neq \mathcal{N}_i(a_{-j})$. Let π_i denote the baseline payoff function of player X_i. The norm-based utility function is given by:

$$u_i(\mathbf{a}) = \pi_i(\mathbf{a}) - k_i \max_{a_{-j} \in \mathcal{L}_{-j}} \max_{m \neq j} \{\pi_m(a_{-j}, \mathcal{N}_j(a_{-j})) - \pi_m(\mathbf{a}), 0\} \tag{1}$$

where $k_i \geqslant 0$ is a parameter specifying X_i's sensitivity to the established norm. While the first maximum operator considers the possibility that a norm might apply to multiple players, the second one ranges over all the players except for the norm violator and specifies the maximum payoff deductions derived from all norm violations.

Although this model can be used to characterize an agent's preference for conformity under the assumption that the empirical and normative expectations conditions are satisfied, the fact that a specific pattern of empirical and normative expectations can *change* the baseline utility to promote conformity to a behavioral rule is left implicit and exogenous. To partially overcome this limitation, Bicchieri and Sontuoso (2017) have proposed extending Bicchieri's framework to dynamic psychological games, a generalization of standard game theory that has been developed precisely to represent motivations that range on beliefs and expectations rather than on actions only (Battigalli & Dufwenberg 2009). As with the original Bicchieri model, the norm-based "psychological" utility of a player is conceived as a linear combination of her material payoff and a norm-based component. However, this latter component is now conceived as an anticipated negative emotion and is a function of a positive difference between the initially *expected* payoff to X_m and the payoff that X_m would get in case of a violation of the behavioral rule. Drawing on the Battigalli and Dufwenberg (2007) concept of *simple guilt*, Bicchieri and Sontuoso end up modeling norm compliance as an aversion to disappointing others' *empirical* expectations. Besides failing to actually take *normative* expectations into account (see Tummolini et al 2013 and Andrighetto, Grieco and Tummolini 2015 for empirical evidence), approaches based on latent psychological motivation fail to explictly capture the kind of conditionality that is required by Bicchieri's own analysis of social norms.

The Bicchieri and Sontuoso model resembles a small tradition of models by economists that analyze implications for equilibrium dynamics of the insertion of fixed (but varying across agents) preferences for conformity with others into individuals' utility functions. The seminal model in this literature is Bernheim (1994), and important theoretical extensions are developed by Brock and Durlauf (2001) and Michaeli and

Spiro (2015, 2017). This strand of theory has been experimentally applied by Andreoni and Bernheim (2009) and Andreoni, Nikiforakis, and Siegenthalier (2017). These authors all refer explicitly to equilibria they derive from preferences for conformity as 'norms', which vary in efficiency. Agents *do* undergo preference shifts in the Bernheim-style models, but this is exogenously imposed rather than endogenously driven by specifically normative dynamics. Consequently, the focus in this tradition has tended to generalise in the direction of the 'herding' literature (e.g. Banerjee 1992, Chamley 2004) that addresses informational dynamics in markets where agents infer asset values from observing others' choices. This has generated some experimental applications (e.g. Duffy & Lafky 2019) where, although the language of 'norms' is featured, they are the object of inquiry only at such an abstract level that any peculiarly normative dynamics disappear from view. Financial asset markets, a prime domain of application for herding theory, are arguably a setting where, because all agents are expected to aim to maximize expected monetary returns, norms in our and Bicchieri's sense, that is, social structures that coordinate descriptive and normative expectations, play no role at all. Notably, Chamley's (2004) advanced textbook on herding includes no index entry for 'norms'.

In political science contexts, attention has frequently focused on cases where disagreements about norms are thought to be relevant to analyzing shifts in policies or coalitions. Norms in such contexts are typically assumed to be standing commitments by subsets of political actors to specific 'ways of doing things', and not merely generalized preferences for conformity. Review of the literature and topics surveyed in Druckman et al (2011) indicates the absence of an experimental literature in political science focused on norms per se, notwithstanding widespread use of Investment / Trust games (Wilson & Eckel 2011). Such experiments are often used to furnish evidence of behavioral sensitivity to hypothesized social preferences, but descriptive and normative expectations are not distinguished from one another, let alone separately estimated so that alignment can be assessed.

Attention to norms is arguably crucial to integrating pioneering work by Kuran (1995) on the dynamics of public opinion, and political responses to these dynamics, with the experimental traditions from economics on which experimenters in political science often draw. Kuran concentrates on cases where what he calls agents' 'intrinsic' utility (meaning utility that is independent of social context, e.g. utility derived from a policy's effect on one's portfolio value) drops out of analysis because agents' preferences and choices concerning social conformity cannot influence it. While not framing his analysis in terms of responses to norms per se, Kuran invites us to focus on causally effective networks of descriptive and normative expectations as social structures. Furthermore, Kuran considers, as we do, norms as drivers of endogenous preference changes at the level of individuals. Other similarly synthetic work by economists that begins to come to grips with social structures as causal mechanisms that influence preferences is Akerlof and Kranton (2010), who consider the complex and important web of relationships between norms and social identities. Here we leave identity as a topic to which the modeling techniques we go on to present might usefully be extended in the future.

In Kuran's basic model, agents face the recurrent problem of deciding whether, and under what social and political circumstances, to express their 'true' private preferences or to instead express preferences that align with 'public opinion', which Kuran identifies with modes of distributions of publicly expressed preferences. This *can* be a problem even when an agent's private preference aligns with public opinion, because the agent might have incentives to appear to be uncommitted or rebellious. But the primary interest concerns cases of misalignment between private and publicly signaled preferences. Kuran argues, with a range of non-hypothetical examples, that it is a pervasive element of the human social and political predicament that people are regularly confronted with choosing trade-offs between their *expressive* utility, derived from exercising and demonstrating their autonomy and self-authenticity, and their *reputational* utility, which derives from the social rewards and sanctions associated with, respectively, conformity and dissent. By 'reputation' Kuran refers not only to an agent's relative valuation by other agents, but to utility that parties to an interaction receive through coordinated expectations, which may be both descriptive and normative. Kuran also recognizes *intrinsic* utility, as characterized above. For example, a wealthy person might benefit from

the abolition of capital gains tax independently of whether she thinks, from a detached point of view, that such abolition is good public policy, and also independently of any social rewards or sanctions that come her way from expressing an opinion that conforms with or diverges from those of others, or is expected by them. Kuran's main analysis sets intrinsic utility derived from outcomes as exogenously fixed, on grounds that in typical political contexts most agents' expressions of preferences have no special influence on what policy choice will in fact prevail. That is, the 'client' to be served by Kuran's primary analysis is not a President or a celebrity activist.

Pervasive tension in real social and political life between expressive and reputational utility gives rise, as Kuran shows, to a range of recurrent patterns. It explains why people often conceal religious beliefs in secular settings, and why secularists disagree among themselves over whether public religious displays should be encouraged or restricted. It explains why there is controversy over whether members of disparaged minorities (e.g. LGBT people) should conceal their identities or give comfort and strength to their fellows by revealing their identities. It explains why politicians arrange anonymous leaks as trial balloons. It explains the very point of secret ballots, blind refereeing, and conducting elite political bargaining behind closed doors. In general, the *preference falsification* that responds to incentives associated with reputational utility blocks straightforward inference from the distribution of publicly expressed preferences to the distribution of true private preferences.

Again, Kuran is most interested in the social dynamical effects of preference falsification. In cases of bimodal public opinion, preference falsification can lead to polarization into extremist camps, if people who express moderate opinions find themselves sanctioned by both sides. People of only slightly less moderate views in either direction have rational incentive to deliver such sanctions, so as to grow the pressure mass of their own faction. In addition, preference falsification often promotes the phenomenon extensively studied in social psychology by Katz and Allport (1931) under the label of 'pluralistic ignorance'(PI). PI is a major theme to which Bicchieri (2006, 2017) applies her analysis of norms when she turns to policy implications. It obtains when a behavioral pattern, regime, or policy that the majority of a population dislikes prevails because no one can observe that their private dispositions or opinions are widely shared. In the case of a norm, as discussed by Bicchieri (2017), PI arises when expectations supporting the norm are maintained and repeatedly confirmed by experience, despite widespread or even majority disapproval of the norm in question, because the extent of the disapproval is invisible or concealed. In Kuran's terms, concern for reputational utility can lead agents to sanction violators of norms that they themselves would secretly prefer to violate, or do in fact violate when they are out of public view. Kuran cites, as a familiar example, gay people in oppressively heteronormative environments engaging in demonstrative homophobia. This pattern can lead to unpredicted and sudden normative lurches in public opinion. An example is the very rapid shift in public opinion in most developed countries in the late 1980s from viewing drunk driving as a reckless but often amusing misdemeanor to viewing it as a breach of social morality deserving criminal prosecution (Lerner 2011).

In the context of our interest in conditional preferences, a particularly interesting feature of public opinion dynamics discussed by Kuran is that preference falsification often leads to preference *revision*, as a person's expressed preferences over time become integrated into the social identity she continuously constructs for herself, and on which others with whom she interacts ground expectations about her actions. This has echoes of Aristotle's view that people become moral agents by becoming habituated to behaving morally. As Kuran stresses, however, this process can apply equally to morally destructive norms, for example the biologically interpreted racism that was socially diffused through most Western societies during the 19th century period of European and American global imperialism (Hanneford 1995), and for decades was normatively established in the largest part of these populations. This is the aspect of dynamic preference influence through networks that most directly corresponds to norm *origination*. Such diffusion and entrenchment typically involves little or no deliberative reflection on the part of most individuals, and in that sense is more accurately modeled as a social process than as a psychological one.

As noted, in Kuran's model agents vary in the weights attached to intrinsic, expressive, and reputational utility in composing their total utility functions. In large-n cases, as explained, intrinsic utility is strategically irrelevant, so we need only attend to variation in binary weightings between expressive and reputational utility. Kuran refers to agents who attach much higher weight to expressive than to reputational utility as 'Activists'. Such agents will tend not to falsify their preferences even when these preferences are socially unpopular and expression of them generates sanctions. They are, then, more likely to violate norms. Types of such Activists include ideologues, religious fundamentalists, moralists, and people whose identities are non-negotiably associated with sectarian lifestyles that cannot be concealed from others. Since Activists limit the extent to which public opinion hides the existence and distribution of falsified private preferences, their presence mitigates against PI, and makes dynamic social norm-reversal more likely. Bicchieri (2017) introduces essentially the same concept, though without Kuran's analytic structure, under the label of 'Trendsetters'. Below we will, in the spirit of fair attribution, call agents with such total-utility composition functions 'Activists / Trendsetters'.

When he models individual agents as 'norm-takers', in the same sense that economists model individuals in markets as price-takers, Kuran abstracts from dynamical interactions between norms as social structures and potential individual preference shifts in response to social influence. But he does not deny that there is such an aspect. Large-scale norms emerge from (and feed back upon) the formation of norms in smaller sub-networks. Such dynamics are important in the laboratory setting of a typical Investment / Trust game experiment. Subjects in such experiments may be expected to bring into the lab various normative expectations that have diffused through their large-scale cultural environments (Binmore 2007). But the setting is novel for most participants, and it is equally clear across the large literature on these experiments that when subjects play multiple rounds they frequently learn to coordinate on expectations that evolve over the course of play. Interesting though such evolution might often be to an experimenter, she might alternatively be mainly concerned to identify subjects' pre-play descriptive and normative expectations by designing tasks the create conflict between available intrinsic utility and hypothesized norms: where subjects must be offered extra monetary incentive to do what is to their intrinsic advantage, this can reveal their beliefs about the existence and influence of a norm. Agents' influence over relevant intrinsic utility in small-n settings thus furnishes a methodological reason to turn to the laboratory even when the ultimate subject of interest, as in Kuran's work and in most contexts that preoccupy political scientists, is norm distribution at the large-n scale.

We also note that Kuran's basis for factoring out intrinsic utility and for assuming that agents are norm-takers, in exemplary applications of his utility model, are only expository idealizations. We can easily conjure examples of politically important situations in which norms are contested in smaller-scale settings where participants are not norm-takers and influence available intrinsic utility. Consider, for example, recent anxieties about the undermining of norms of professionalism, technocratic control, and responsiveness to scientific consensus in US Government agencies under pressure from an anti-bureaucratic and anti-expert presidential administration. Some agencies are reported as resisting this pressure much more successfully, or at least for longer, than others (Lewis 2018). It is unlikely that officials engaged in such struggles, with acute self-awareness, perceive themselves to be bereft of individual influence on the outcome, even if they are not optimistic about their capacity to successfully hold out indefinitely. We might expect that in this sort of setting, considerations of expressive and reputational utility exert strong influence along with the intrinsic utility called into play by individual influence. A great deal of attention in political science is devoted to smaller-scale institutional settings of this kind.

A natural general question to ask about such instances is: what are the characteristics of a network that makes its norms more or less fragile under competitive pressure from alternative or (as in the example above) subversive norms?

Progress toward an adequate formal and operational model mandates, then, the adoption of a new approach enabling the possibility for agents to endogenously modulate their preferences in response to discov-

eries about population-scale distributions of empirical and normative expectations, and not just to agents' pre-standing preferences concerning possible actions of others with whom they individually happen to interact. We seek a model according to which agents do not just *express* norms through their choices of strategies and actions, but *recalibrate* their own preferences upon encountering norms as social structures. We follow a course of conservatism in one respect, however. We worried earlier about the degrees of freedom allowed in state-dependent utility models by the absence of a principled general ontology of exogenous states of nature. Furthermore, we follow Binmore (2010) in objecting that the social preference approach is unduly liberal in allowing any equilibria in games used as models to be rationalized by freely conjecturing specific and idiosyncratic arguments in utility functions. There would be no obvious methodological gain to be expected from swapping degrees of freedom in state-dependent and social preference theories for abandonment of general constraints on specifications of utility functions and solution concepts. Therefore, in our analysis to follow, we adopt the following practices. First, we apply Kuran's version of norm-sensitive utility rather than Bicchieri's, on grounds that the former is more general. Second, we deploy an extension of game theory, conditional game theory (Stirling 2012, 2016) that does not *refine* standard solution concepts by imposing special restrictions on strategic rationality as in theoretical work by Kreps (1990) and Bicchieri (1993).

3 Requirements for identifying and estimating models of norms in empirical data

The experimental literature on norms is extensive, and includes both laboratory and field experiments. No systematic cross-disciplinary survey as yet appears to exist, which arguably reflects the lack of conceptual unification to which we seek to make a partial contribution. Attempting such a survey would be beyond the scope of the chapter, and tangential to its purpose, which is to provide a formal account of the strategic dynamics of norms that is consistent with the philosophical analysis provided by Bicchieri (2006, 2017), and can form the basis for estimating experimental data.

Part of the motivation for this challenge is that the formal approach suggested by Bicchieri and her various co-authors for estimating their own experimental data, as reported and analysed in Bicchieri and Xiao (2009), Xiao and Bicchieri (2010), Bicchieri and Chavez (2010), Bicchieri, Xiao, and Muldoon (2011), Bicchieri and Chavez (2013) and Bicchieri, Lindemans, and Jiang (2014), models social norm compliance as the optimization of a norm-based utility function in which, as discussed in Section 2, the losses of others figure as an explicit argument rather than allowing preferences for conformity to arise endogenously through strategic interactions. In consequence, analyses of the experiments consist in identifying specific behavioral frequencies and characteristics rather than estimating parameters of a theory of norms. Similarly, the authors of the rich vein of experimental evidence for the influence of experience of market-like institutions and transactions (along with adherence to Christianity or Islam) on normatively governed strategic behavior in small-scale societies (Henrich et al 2004; Ensminger Henrich 2014) assume that norms reflect individuals' social preferences, but offer no comments on whether norms are social structures that influence individual social preferences, or are simply emergent aggregate descriptions of the social preferences themselves.[2] Bicchieri (2006, 2017), at least, offers the basis of a theoretical specification of a social norm, but then does not bring it clearly to bear on analysis of her own (and co-authors') experimental observations. Our main aim in the chapter is to close this methodological gap.

Before we take up this task, however, we point to some limitations in the experimental methods used in the studies by Bicchieri and co-authors cited above. We concentrate on these studies both because they are motivated by the philosophical conception of norms we aim to develop, and because in one crucial respect they are improvements on methods generally used in experiments on norms that have been conducted by

[2]Henrich et al (2004) at one point (p. 376) suggest that norms at least sometimes reflect institutionalised practice, but offer no comment on the dynamics of relationships between such institutions and individual preferences.

psychologists and political scientists. The respect in question is that Bicchieri and her co-authors appreciate the importance of eliciting both first-order empirical expectations of subjects about what other subjects *will* do in games, and their second-order normative beliefs about what players are expected or *should* do, using *salient incentives*. This is generally essential to valid inference in all behavioral experiments (Harrison 2014), but is particularly crucial in eliciting beliefs about social and personal values, where even subjects aiming to be sincere may be motivated by interest in self-signalling of virtue or in minimising dissonance between social expectations and their self-conceptions (Bicchieri 2017). We indicate improvements in belief elicitation methods that are important for the sake of structural modeling using general theory. But where the experiments of Bicchieri and her co-authors are concerned, it is gratifying not to need to argue for introducing incentives in the first place.

Many experiments reported by Bicchieri are based on variants of the 'Investment' or 'Trust' game as discussed in Section 2. In the lab, subjects may play the game sequentially, or using 'the strategy method' in which players simultaneously submit their decisions over discretized sets of transfers at every possible decision node for one or both roles. As noted previously, players can achieve Pareto efficient outcomes if they operate normative expectations, under one or another interpretation of 'norm'. The least determinative of such expectations is simply that non-zero levels of trust and reciprocity are expected, in which case the game is a pure coordination game. An influential literature using the framework of psychological game theory purports to test hypotheses that players' norm-based utility functions include some form of social preference for egalitarian outcomes or proportional reciprocity (Azar 2019), but do not elicit incentivised reports of subjects' beliefs about the extent to which they expect others to share such preferences. The form of interpretation of evidence we would need for consistency with Bicchieri's analysis of norms is that players might share consistent descriptive and normative expectations that, in a class of interactions that includes the Investment / Trust game setup, they use to regulate their behavior, and may normatively endorse, such as the 'Equality' or 'Equity' norms mentioned in Section 2. Investment / Trust games conducted in the laboratory can provide evidence of the existence of norms governing communities from which subjects are drawn just in case players' first- and second-order empirical and normative beliefs are elicited for comparison with their chosen strategies (Bicchieri 2017).

Below we describe procedures that can straightforwardly be applied to incentivise subjects' first-order descriptive beliefs about the frequency of a norm in a population with which they experimentally interact. By contrast, it is not possible to incentivize subjects' first-order normative beliefs, since in the absence of some currently unavailable (and perhaps in principle incoherent) neuropsychological test, there are no independent measures against which such reports could be compared. Both second-order descriptive and normative beliefs, on the other hand, can be incentivized by asking subjects to predict reports of other subjects and rewarding such predictions based on their accuracy. Bicchieri and her co-authors have generally followed this practice. Bicchieri, Xiao and Muldoon (2011) incentivized second-order descriptive beliefs but not second-order normative beliefs. Bicchieri and Xiao (2009), Xiao and Bicchieri (2010), and Bicchieri and Chavez (2010, 2013) incentivized both types of beliefs. These experiments have thus represented progress in the direction of best practice. However, the incentivization methods have all involved serious limitations with respect to the goal of identifying and estimating structural models of utility conditioned on beliefs about social structures.

A first, straightforward, limitation is that all of the experiments above used reward magnitudes for second-order belief elicitation that were arguably too small to be effectively salient. University student subjects were paid one US Dollar for correct point predictions, and otherwise zero. Of course the way to overcome this limitation is simple and obvious.

A more interesting problem is that in all of the experiments above, what are elicited are only distributions of subjects' point estimates of modal beliefs. This provides no basis for estimation of any individual subject's descriptive or normative expectations as explanatory factors for choices. In realistic settings, what the analyst who aims to predict normatively sensitive behavior needs to know are distributions of an individual

agent's expectations in social contexts, and the relative *confidence* with which an individual agent expects that responses will be norm-governed, since, on Bicchieri's analysis, she will conform her own choices to norms only to the extent that she believes that those with whom she interacts will do so.

Methods for eliciting beliefs as degrees of confidence have been developed. For example, Harrison et al (2017) ask each subject to distribute tokens over a set of 'bins' that range across the possible realization values of variables about which the experimenter wants to discover their beliefs. Subjects are rewarded according to a proper scoring rule, specifically the quadratic scoring rule (QSR), which maps probability mass functions onto payoffs distributed around the realized outcome (Matheson & Winkler 1976). Subjects using the interface presented in Harrison et al (2017) need not learn to represent the rule explicitly. As they operate a slider that controls token allocations over the bins, they are shown money rewards they can expect if the realization corresponds to the center of the probability mass of the allocation they have tentatively selected. Experience in the laboratory is that subjects quickly learn to manipulate allocations fluidly and confidently.

This elicitation in itself is not sufficient for estimating subjects' confidence in their beliefs unless it is known that they are risk-neutral subjective expected utility maximizers. In the Trust game experiments of Bicchieri and co-authors, this is a maintained a priori assumption. However, although many people in laboratory experiments designed to elicit risk preferences through incentivized choices over pairs of lotteries do choose consistently with Expected Utility Theory (EUT), Harrison and Ross (2016) report that majorities in most samples exhibit *rank dependence*. Rank-Dependent Utility Theory (RDU) (Quiggin 1982) describes a family of specifications of utility that nest EUT but allow for subjective decision weights on lottery outcomes, indicating that subjects display relative pessimism or optimism with respect to outcomes depending on their ranking of these outcomes from best to worst. Furthermore, most people, at least in risky decisions involving money, are moderately risk averse (Harrison & Rutström 2008).

A decision by a person to follow a norm, on Bicchieri's analysis, necessarily involves risk, since according to that analysis she will regard this decision as correct only if those with whom she interacts will do likewise, and that they will do so is typically an uncertain conjecture. Clearly, in the Investment / Trust game experimental setting both Investors and Trustees make risky choices, and the game would be of little interest otherwise. Thus there are two reasons for eliciting risk preferences: using the observed lottery choices to identify and estimate structural utility models (i.e., allowing for RDU) at the level of the individual, and incorporating these models in analysis, when investigating norms using Trust games. First, assuming as Bicchieri and co-authors do, that people are risk-neutral expected utility maximizers will typically involve maintaining a counterfactual, leading to incorrect characterizations and predictions (Chetty et al 2020). Second, estimating a subject's rank-dependent adjustment and confidence in her beliefs, based on elicitation using the QSR of her priors on the distribution of probabilities of outcomes, depends on independent elicitation of her risk preferences. This is typically done using lottery-choice experiments (Harrison & Rutström 2008).

Much of the existing experimental literature on norms involves use of repeated games that are designed to allow for observation of learning by subjects, typically interpreted as learning about norms, over the course of play. This makes good sense where, as is typically the case, the laboratory design confronts subjects with what is expected to be a novel situation for them. In other cases, however, where the point of an experiment is to identify norms that players bring with them into the lab, one-shot designs are more appropriate. In these cases, the experimenter who lacks empirical access to convergence on equilibria over time might yet want to be able to estimate what *would* happen dynamically in the limit, particularly if she is interested in the welfare implications of the normative structures she finds, but expects, as Bicchieri (2017) emphasises, that these should often take account of scope for norms to influence preferences through interactive learning. Following standard methodology, she might do this by simulating Bayesian learning over a hypothetical sequence of play. This will tell her about expected equilibria in beliefs and actions. However, it is not clear in advance of further analysis of the operationalisation of the concept of a norm how she

might simulate the endogenous evolution of norms themselves, unless norms are identified with individuals' beliefs. But it is precisely the advantage of Bicchieri's analysis that although agents' decisions over whether to follow norms are based on their individual expectations about others' actions (thus requiring that these be estimated in empirical applications), the norms themselves are functions of *networks* of expectations. The analyst who aims to understand how a conjectured norm might lead agents to potentially improve their welfare by adapting their preferences to accommodate the norm needs a method that allows her to rigorously idealize expectations without having to project such idealization onto subjects whose choices might involve erroneous beliefs.

What is needed, then, for adequate formalization of Bicchieri's philosophical analysis of norms if the theory in question is to be used to specify intended models of laboratory data, is that it allow scope for (i) varying degrees of risk aversion, (ii) rank-dependent utility, and (iii) varying full distributions of subjective beliefs about probabilities of outcomes, but without (iv) depending on empirical observation of learning in real time. Our theory construction below respects these desiderata.

4 Conditional Game Theory

4.1 Conditional Game Preference and Solution Concepts

In this section we explain the concept of conditionality as it is formally constructed by conditional game theory (CGT). We will subsequently demonstrate the power of this construction to simultaneously express the idea of conditional preference we attributed to Bicchieri (2006, 2017), and identify dynamic normative influence in observed choice data. The reader who would like to see more of the formal background theory is referred to the Appendix, and to Stirling (2012, 2016).

The essential components of a game are (a) a set of n agents, denoted $\mathbf{X} = \{X_1, \ldots, X_n\}$, each with its *action set*, denoted $\mathcal{A}_i = \{x_{i1}, \ldots, x_{iM_i}\}$, $i = 1, \ldots, n$; (b) the set of *outcomes* as a function of the *action profiles*, denoted $\mathbf{a} = (a_1, \ldots, a_n) \in \boldsymbol{\mathcal{A}} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$; (c) a *preference concept* for each X_i that specifies X_i's metric by which the elements of $\boldsymbol{\mathcal{A}}$ set are evaluated; and (d) a *solution concept* that defines equilibria. Thus, a game analysis reconciles the fact that the consequence to each agent depends on the choices of all agents, but each agent has control only over their own actions. Under standard noncooperative game theory, each X_i defines her preferences via a utility function u_i: $\boldsymbol{\mathcal{A}} \rightarrow \mathbb{R}$, and solutions are defined as strategically rational choices—that is, actions drawn from $\mathcal{A}_i$ that maximize X_i's welfare under the expectation that others will do likewise (e.g., Nash equilibria, dominance).

The major difference between a standard noncooperative game and a conditional game involves the preference concept. Standard game theory defines preferences *categorically*, by specifying fixed, immutable, and unconditional utility for each player. This structure requires players to respond to the expected actions of all others as specified by a selected equilibrium. The key innovation of CGT is that, whereas categorical preferences are declarative statements of unconditional preference, CGT allows agents to modulate their preferences by responding locally to the social distribution of preferences, not just to the expected actions of others. A standard noncooperative game can be re-expressed as a conditional game by reinterpreting the standard utility as a conditional utility of the form $v_{i|-i}$: $\mathcal{A}_i|\mathcal{A}_{-i} \rightarrow \mathbb{R}$ (where, similar to the notation introduced in Section 2, the expression $v_{i|-i}$: $\mathcal{A}_i|\mathcal{A}_{-i}$ denotes the conditional utility for X_i given the possible actions of others) such that $v_{i|-i}(a_i|a_{-i}) = u_i(a_i, a_{-i})$. Although the conditioning is with respect to the possible *actions*, a_{-i}, of the others, that is not the same as conditioning on the *preferences* of the others. This distinction has been recognized by Manski (2000), who argues that

> A more general class of interactions permits the preferences, expectations, and constraints of one agent to affect the preferences, expectations, and constraints of another agent in ways that are not mediated through actions. *It is one thing to say that my preferences depend on your*

> *actions, and another to say that my preferences depend on your preferences* [emphasis added] (pp. 120-121).

Explicit allowance for conditioning preferences provides a framework with which to model complex social interrelationships. In particular, it provides an explicit apparatus for modeling normative behavior, as Bicchieri understands it, that is consistent with her philosophical analysis.

In CGT, preferences are expressed as hypothetical propositions, with the preferences of others as antecedents and actions of the reference agent as consequents. This structure exactly parallels the logical structure of Bayesian probability theory as applied to scientific reasoning, namely, conditionalization - the process of determining the cumulative influence of acquired evidence. The best known application of Bayesian probability is as a device for modeling epistemological theories. The cumulative process of belief revision in response to updated evidence is expressed as a Bayesian network with vertices as random variables and edges as conditional probability mass functions that define the statistical relationships among the random variables. These conditional relationships are combined via the chain rule to generate a joint probability distribution that incorporates all of the interrelationships that exist among the random variables. Once defined, unconditional belief orderings for each random variable can be deduced by marginalization.

CGT appropriates the Bayesian network syntax to model a praxeological theory of preference and preference recalibration based on social influence. The general model defines a three-phase *meso-to-macro-to-micro* process comprising socialization, diffusion, and deduction, as illustrated in Figure 4.1. Socialization (the meso, or intermediate-level phase) is achieved by expressing preferences as *conditional utility mass functions* that modulate individual preferences as functions of the preferences of those who socially influence them. Diffusion (the macro, or global-level phase) is represented by modeling the social influence network with the syntax of a Bayesian network with agents as vertices and conditional payoffs as edges. As social influence diffuses through the network via these linkages, nascent social interrelationships emerge, thereby generating a *coordination function* that is isomorphic to a joint probability mass function. Deduction (the micro, or local-level phase) is then achieved by marginalization, yielding *individual coordinated payoff functions* that can be analysed using standard game-theoretic solution concepts.

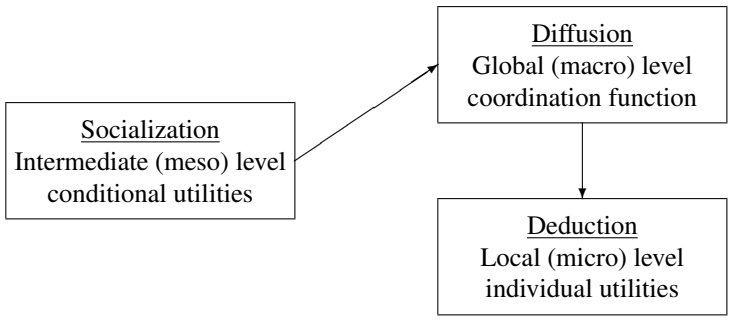


Figure 4.1: The socialization, diffusion, deduction process.

Diffusion is an iterative dynamic process of complex multidirectional and reciprocal mindshaping (Zawidzki 2014; Ross & Stirling 2020), where each agent responds to social influence exerted by her neighbors and, in turn exerts social influence on her neighbors, who again exert influence on her, and so forth. There are two possible ways for such iterative behavioral influence to play out. Under some circumstances it can result in non-terminating but repeating oscillations, in which case nothing gets resolved. But for a range of social situations modeled as games in the literature (Stirling 2016) diffusion results in convergence to unconditional or *steady-state* utilities for each player, where each player possesses totally ordered preferences that form the basis for standard equilibrium analysis. Expressing diffusion with the probability syntax

is restricted by two important conditions. First, we require that diffusion be *coherent*, that is, each agent must have "a seat at the table" in the sense that her utility function influences resolution. In other words, no agent may be subjugated (i.e., disenfranchised) by her neighbors. Of course, real human communities often oppress individuals and sub-communities. Our restriction is technical: in CGT, we would model such exclusion as a relationship between networks rather than as a relationship within a single network. Second, the process must *converge*, in that it results in unambiguous criteria that enable all agents to make coherent choices. Stirling (2012, 2016) establishes that both of these conditions can be satisfied if and only if the diffusion process complies with the probability axioms. Specifically, by requiring the conditional utilities to conform to the syntax of probability mass functions and combining them according to the rules of probability theory (e.g., conditionalization, the chain rule, Bayes's rule), we may invoke two fundamental theoretical results from probability theory: the Dutch book theorem and the Markov chain convergence theorem. A Dutch book is a gambling scenario that results in a sure loss, and the Dutch book theorem establishes that a sure loss is impossible if and only if the gambler's beliefs and behavior comply with the probability axioms. In our context, subjugation is isomorphic to a sure loss (Stirling 2016), and it follows that such a condition is impossible if and only if all agents' preferences and behavior comply with the probability axioms. Furthermore, by viewing the diffusion process as a Markov chain, the Markov convergence theorem can be applied to establish that each agent's preferences converge to a unique utility function that takes into account all of the social relationships that are generated by the conditional utilities.

4.2 The Investment / Trust Game

We will illustrate the application of CGT to the theory of norms by reference to the Investment / Trust game. As discussed, this game was originally introduced as a two-agent game between an Investor, who possesses an endowment $\mathcal{E}$ and a Trustee who manages the investment. The Investor sends $\sigma\mathcal{E}$, with $0 < \sigma \leqslant 1$, to the Trustee and retains $(1-\sigma)\mathcal{E}$. The standard model is that this investment is exogenously (e.g., by an experimenter or by a market process) tripled in value to $3\sigma\mathcal{E}$, resulting in combined non-integrated wealth of the two players of $(1+2\sigma)\mathcal{E}$. The Trustee then returns a portion of her holdings to the Investor with the amount returned depending on the normative posture of the Trustee. For purposes of an illustrative example throughout the chapter, we consider two possible norms that might regulate the choices of players:

N_1**: Equality:** A fair return is one that equalises the final (non-integrated) positions of the players. Under this norm, the payoff to both Investor and Trustee is $(1+2\sigma)\mathcal{E}/2$.

N_2**: Equity:** A fair return is one that is proportional to the share of the endowment that was invested. The Trustee returns the same fraction of the multiplied outcome to the Investor, that is, she returns $3\sigma^2\mathcal{E}$. Thus, the payoff to the Investor is $(1-\sigma+3\sigma^2)\mathcal{E}$ and the payoff to the Trustee is $3(\sigma-\sigma^2)\mathcal{E}$.

The hypothetical setting for our analysis of the Investment / Trust game throughout the paper will be a scenario in which there are two initial communities, where one community begins with prevailing expectations that the game is played under governance of the Equality norm, and the other community begins with prevailing expectations that the game is played under governance of the Equity norm. We will investigate various scenarios under which these communities might encounter one another, in the sense of an agent drawn from one community finding herself playing Investment / Trust against an agent drawn from the other community. This will allow us to examine patterns by which normative variation influences individuals' preferences, as revealed by probabilities of choices of actions. Because players also interact with members drawn from their original normative communities, each agent's preferences are subject to two channels of normative influence: direct influence from expectations concerning play against an agent governed by a 'foreign' norm, and indirect influence of this exposure through its affect on the expected play of 'domestic' game partners. Thus our setup will reflect Bicchieri's core idea that norms are networks of expectations on which preferences are conditional.

In standard versions of the Investment / Trust game in the literature, the Investor is free to decide whether to transfer any amount > 0 to the Trustee. Because in this setting the Trustee can take no action that expresses her normative preferences, for simplicity and transparency of the mechanism of interest, we exclude pure self-interest from the set of available norms. In a laboratory setting this unrealistic restriction would need to be abandoned. We will assume throughout that players' choices are elicited by a 'strategy method', that is, that they choose an action for every node in the extensive form of the game, under the assumption that a random process will allocate them to Investor and Trustee roles. Thus, where players coordinate on one of the two norms above, the expected monetary value *I* of the game is the mean of the return to the Investor and the Trustee conditional on the norm:

$$\begin{aligned} I_{\mathsf{N}_1} &= (1+2\sigma)\mathcal{E}/2 \\ T_{\mathsf{N}_1} &= (1+2\sigma)\mathcal{E}/2\,, \end{aligned} \tag{2}$$

and

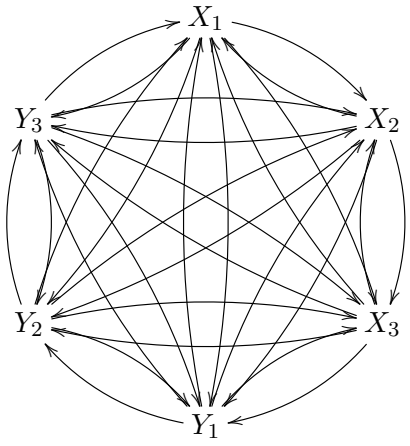
$$\begin{aligned} I_{\mathsf{N}_2} &= (1-\sigma+3\sigma^2)\mathcal{E} \\ T_{\mathsf{N}_2} &= 3(\sigma-\sigma^2)\mathcal{E}\,. \end{aligned} \tag{3}$$

Consider a bilateral transaction between agents Z_i and Z_j. To begin with, suppose that $\mu(I) = I$. Let I_{N_j} be the utility Z_i receives from Z_j who abides by norm $\mathsf{N}_j \in \{\mathsf{N}_1, \mathsf{N}_2\}$, and let T_{N_i} be the utility that Z_i retains. By symmetry, Z_j receives I_{N_i} from Z_i and retains T_{N_j}. This situation may be represented by the network graph

$$Z_i \underset{\tilde{u}_{Z_i|Z_j}}{\overset{\tilde{u}_{Z_j|Z_i}}{\rightleftarrows}} Z_j \tag{4}$$

where $\tilde{u}_{Z_j|Z_i}$ and $\tilde{u}_{Z_i|Z_j}$ are conditional utilities that express the influence that they exert on each other.

We aim to model situations under which groups of agents governed by different norms encounter one another. Therefore, consider a six-agent graph showing the interconnection of two three-agent subnetworks, $\{X_1, X_2, X_3\}$ and $\{Y_1, Y_2, Y_3\}$, of the form



(5)

where agents $\{X_1, X_2, X_3\}$ begin with descriptive expectations that others play as per the Equality norm, and agents $\{Y_1, Y_2, Y_3\}$ begin with descriptive expectations that others play as per the Equity norm. In each conjectured play of the game, each agent chooses from the norm set $\mathcal{A} = \{\mathsf{N}_1, \mathsf{N}_2\} = \{\text{Equality}, \text{Equity}\}$, and each defines her conditional payoff $p_{i|jklmn}$ as a mapping from $\mathcal{A}$ given the conjectures of the other

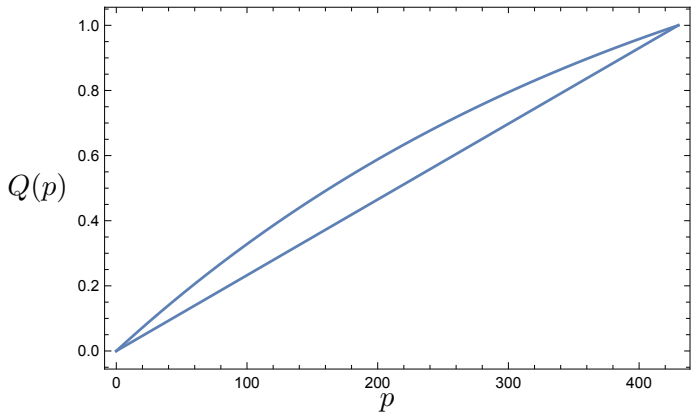


Figure 2: Payoff function without normative expectations.

agents, that is, $p_{i|jklmn}$: $\mathcal{A}|\,\mathcal{A}^5 \rightarrow \mathbb{R}$ with clock-wise indexing convention

$$i|jklmn \in \{X_1|X_2X_3Y_1Y_2Y_3,\ X_2|X_3Y_1Y_2Y_3X_1,\ X_3|Y_1Y_2Y_3X_1X_2,\\ Y_1|Y_2Y_3X_1X_2X_3,\ Y_2|Y_3X_1X_2X_3Y_1,\ Y_3|X_1X_2X_3Y_1Y_2\}\,. \quad (6)$$

Let N_i denote the norm conjecture for the conditioned agent (the agent on the left side of the conditioning symbol "|"), let $\mathbf{N}_{jklmn} = (N_j, N_k, N_l, N_m, N_n)$, denote the norm conjectures for the conditioning agents, and let $I_{N_i}, I_{N_i}, T_{N_i}, T_{N_i}, i, j, k, l, m, n \in \{X_1, X_2, X_3, Y_1, Y_2, Y_3\}$ denote the payoffs for the conditioning agents. The general form of the payoff function is

$$p_{i|jklmn}(N_i|\mathbf{N}_{jklmn}) = I_{N_j} + I_{N_k} + I_{N_l} + I_{N_m} + I_{N_n} + 5T_{N_i}\,, \quad (7)$$

where the utilities are additive if each agent plays an Investment / Trust game with every other player and accumulates the results.

The values expressed by (7) are in monetary units, which must be normalized to conform to the syntax of probability theory. It is thus convenient to now discharge the assumption that $\mu(I) = I$, which we must do in any case to bring the modeling within the rubric of economic theory. We first map the values to the unit interval via a concave transformation over $[0, p_{max}]$, where p_{max} is the largest utility obtainable over all of the arguments of $p_{i|jklmn}$, yielding $p_{max} = 5\max\{I_{\mathsf{N}_1}, I_{\mathsf{N}_2}\} + 5\max\{T_{\mathsf{N}_1}, T_{\mathsf{N}_2}\}$. For purposes that will emerge in due course, a suitable concave mapping is

$$Q(p) = \frac{1 - \exp(p/p_{max})}{1 - \exp(-1)}\,, \quad (8)$$

as displayed in Figure 2. Let

$$\check{p}_{i|jklmn}\big(N_i|\mathbf{N}_{jklmn}\big) = Q\big[p_{i|jklmn}\big(N_i|\mathbf{N}_{jklmn}\big)\big]\,. \quad (9)$$

We then normalize these utilities to become mass functions, denoted $\tilde{v}_{i|jklmn}$, via the transform

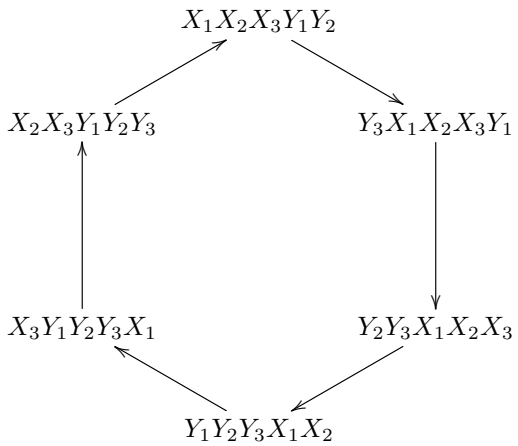
$$\hat{p}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_{jklmn}) = \frac{\check{p}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_{jklmn})}{\check{p}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_i) + \check{p}_{i|jklmn}(\neg\mathsf{N}_i|\mathbf{N}_i)} \quad (10)$$

for $i = 1, 2$, where $\neg\mathsf{N}_i$ is the alternative to N_i.

Standard Bayesian network theory applies only to acyclic (i.e. hierarchical) influence relationships, where influence propagates unidirectionally, and independently specified reciprocal relationships are prohibited (i.e. Bayes's rule must be satisfied). To represent normative influence through conditionalization,

however, reciprocal relationships are indispensable, and the corresponding conditional payoffs must be independently specifiable. The concept of dynamic exchanges amongst individuals is fundamental to mind-shaping as discussed by Zawidzki (2013), that is, the processes by which individuals engineer social environments through imitation, pedagogy, conformity to norms, and coordinated narrative self-constitution, in ways that influence others to modify their beliefs and preferences. To deal with networks of the form in (5), we must generalize beyond hierarchical network structures and accommodate networks with cycles. We begin by recognizing that it is not the concept of reciprocity that is prohibited by Bayes's rule. Rather it is *simultaneous* reciprocity that is problematic. But reciprocity of the type we are considering requires time-dependent exchanges. Consider the two-agent network (4). Z_j's preferences influence Z_i's preferences, which then influence Z_j's preferences, which again influence Z_i's preferences, and so on, ad infinitum. The critical question of interest is whether this sequence of exchanges oscillates indefinitely or converges to a unique limit where each agent is assigned an unconditional utility defined over her action set. The key mathematical tool for this investigation is the Markov chain convergence theorem. In its conventional probabilistic application, this theorem establishes necessary and sufficient conditions for the joint distribution of a set of time-evolving discrete random variables (termed a Markov chain) to achieve a stationary distribution. Because the syntactical structure of a conditional game satisfies the mathematical conditions for the application of the Markov chain convergence theorem, we can apply the theory to compute steady-state (i.e., stationary) utilities.[3]

Our task, therefore, is to model the dynamic relationship between the utility of the normative profiles of the agents as time evolves. This task is particularly challenging due to the complex interrelationships between agents as expressed by the graph displayed in (5). Fortunately, however, a graph of a network is *not* the network; it is only a representation of it, and graphical representations of a network are not unique. To be useful, however, two representations of a network must be *Markov equivalent*, meaning that the conditionalization properties of the two graphs are identical. In particular, we are interested in defining a Markov equivalent graph that converts a graph whose vertices are single agents and whose edges are multidirectional linkages (i.e., (5)) into a graph whose vertices comprise multiple agents and whose edges are unidirectional linkages. As is established in the Appendix, such a Markov equivalent network is



(11)

where the edges linking five-agent subnetworks to five-agent subnetworks are transition matrices $T_{ijklm|jklmn}$,

[3] An important technical property for the application of the Markov chain convergence theorem is that the conditional utilities must satisfy the Markov property, which means that the conditional utility at a given time depends only on the state of the network at the immediately previous cycle.

as defined in the Appendix A.4, for

$$\begin{aligned} ijklm|jklmn \in \{&X_1X_2X_3Y_1Y_2|X_2X_3Y_1Y_2Y_3, X_2X_3Y_1Y_2Y_3|X_3Y_1Y_2Y_3X_1, \\ &X_3Y_1Y_2Y_3X_1|Y_1Y_2Y_3X_1X_2, Y_1Y_2Y_3X_1X_2|Y_2Y_3X_1X_2X_3, \\ &Y_2Y_3X_1X_2X_3|Y_3X_1X_2X_3Y_1, Y_3X_1X_2X_3Y_1|X_1X_2X_3Y_1Y_2\}\,, \end{aligned} \tag{12}$$

and where the entries in these matrices are composed of the conditional utilities $\hat{p}_{i|jklmn}$. The Markov chain convergence theorem then establishes that the steady-state coordination functions for each five-agent subnetwork are the eigenvectors corresponding to the unique unity eigenvalues of the closed-loop transition matrices formed as

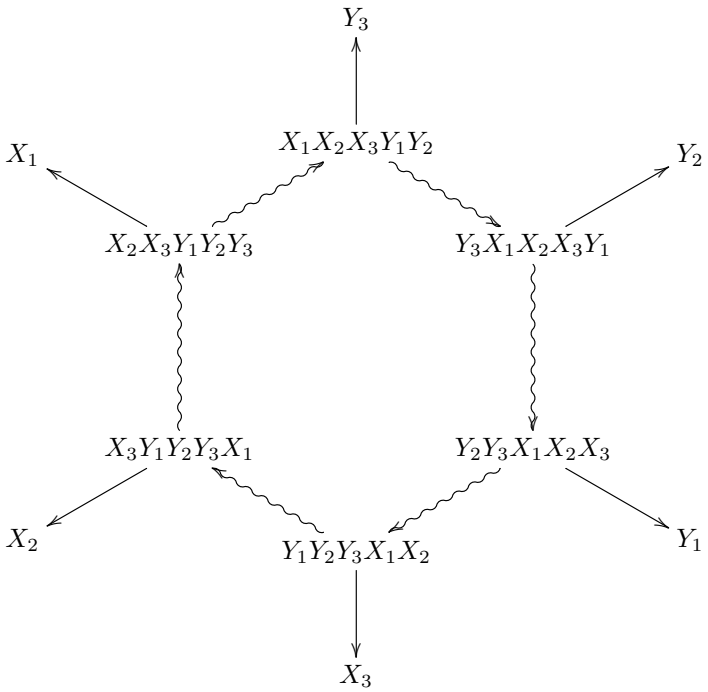
$$T_{ijklm} = T_{ijklm|jklmn}T_{jklmn|klmni}T_{klmni|lmnij}T_{lmnij|mnijk}T_{mnijk|nijkl}T_{nijkl|ijklm}\,, \tag{13}$$

yielding, in matrix format, $\overline{\mathbf{w}}_{ijklm}, \overline{\mathbf{w}}_{jklmn}, \overline{\mathbf{w}}_{klmni}, \overline{\mathbf{w}}_{lmnij}, \overline{\mathbf{w}}_{mnijk}, \overline{\mathbf{w}}_{niijkl}$, as defined by (A.18).[4] We emphasize that convergence is independent of the initial state.

The steady-state network is illustrated in (15), where the edges denoted by $\rightsquigarrow$ are *dormant* — they still exist but are inactive once steady-state (i.e., convergence) is achieved. Finally, the individual coordinated utilities are obtained as

$$\overline{\mathbf{w}}_i = T_{i|jklmn}\overline{\mathbf{w}}_{jklmn}\,, \tag{14}$$

where the transition matrices $T_{i|jklmn}$ are composed using the conditional utilities $\hat{p}_{i|jklmn}$ as described in Appendix A.4.



(15)

We have now shown how to use the mechanism of conditionalization to represent influence on beliefs, which might in principle be normative beliefs, in the setting of the Investment / Trust game with two sets of competing norms we have simply stipulated as such. However, we have yet to introduce any machinery for representing normative expectations. Thus what has been developed so far has not yet formally reconstructed

[4]Notice that it is not necessary to compute the limit as $t \to \infty$ in (A.18). Once the closed-loop transition matrices are defined, the steady-state vectors are immediately available upon the calculation of the relevant eigenvectors.

Bicchieri's analysis. Conditionalization as developed to this point operates only over potential utility gains from adjustments of descriptive expectations when the X_n and Y_n networks are combined. But if we can formally capture Kuran's cases of normative change, which depends on dynamics of expectations, then it follows that we will have shown that our formal operationalization of conditionality is isomorphic to Bicchieri's informal idea of it.

5 Normative Dynamics

5.1 Modeling Conflicting Norms in the Investment / Trust Game

As discussed above, Kuran (1995) models an agent's *total utility* from a transaction as being additively composed of her *intrinsic* utility *I*, her *reputational* utility R, and her *expressive* utility E. Adapting this framework to the Investment / Trust game, we will use I to denote the monetary value of payoffs, which will be transformed into utilities. From Subsection 4.3, agents' baseline governing norms will be drawn from the set $\mathcal{A} = \{\mathsf{N}_1, \mathsf{N}_2\} = \{\text{Equality}, \text{Equity}\}$. For simplicity, we will initially suppose that R and E are 'all or nothing', that is, that $R \in \{0, 1\}$ and $E \in \{0, 1\}$, where $R = 1$ when a player chooses the action expected by her partner, and 0 otherwise, and $E = 1$ when a player chooses the action mandated by her private preference for Equality or Equity, and 0 otherwise. Then the general form of a player's 'Kuran utility' for the game will be

$$\omega(I, R, E) = \omega(\alpha + I + \beta R + \delta E) \tag{16}$$

which, expressed as a conditional payoff as discussed in Section 4.2, (see (7)), becomes

$$\begin{aligned}\omega_{i|jklmn}(N_i|N_j, N_k, N_l, N_m, N_n) = \alpha + I_{N_j} + I_{N_k} + I_{N_l} + I_{N_m} + I_{N_n} + 5T_{N_i} \\ + \beta\big(R_i(N_i|N_j) + R_i(N_i|N_k) + R_i(N_i|N_l) + R_i(N_i|N_m) + R_i(N_i|N_n)\big) \\ + \delta\big(E_i(N_i|N_j) + E_i(N_i|N_k) + E_i(N_i|N_l) + E_i(N_i|N_m) + E_i(N_i|N_n)\big)\,,\end{aligned} \tag{17}$$

where α is the player's baseline monetary assets, and β and δ are independent parameters that determine the shadow prices, in the currency of *I*, of R and E respectively. The expression

$$I_{N_j} + I_{N_k} + I_{N_l} + I_{N_m} + I_{N_n} + 5T_{N_i} \tag{18}$$

is the intrinsic component,

$$\beta\big(R_i(N_i|N_j) + R_i(N_i|N_k) + R_i(N_i|N_l) + R_i(N_i|N_m) + R_i(N_i|N_n)\big) \tag{19}$$

is the reputational component, and

$$\delta\big(E_i(N_i|N_j) + E_i(N_i|N_k) + E_i(N_i|N_l) + E_i(N_i|N_m) + E_i(N_i|N_n)\big) \tag{20}$$

is the expressive component.

Following the procedures introduced in Section 4.2, $\omega_{i|jklmn}$ must be transformed via (9) and (10), resulting in

$$\breve{\omega}_{i|jklmn} = \frac{1 - \exp(\omega_{i|jklmn}/\omega_{max})}{1 - \exp(-1)} \tag{21}$$

with $\omega_{max} = 5\max\{I_{\mathsf{N}_1}, I_{\mathsf{N}_2}\} + 5\max\{T_{\mathsf{N}_1}, T_{\mathsf{N}_2}\} + 5\beta + 5\delta$ and

$$\hat{\omega}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_{jklmn}) = \frac{\breve{\omega}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_{jklmn})}{\breve{\omega}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_i) + \breve{\omega}_{i|jklmn}(\neg\mathsf{N}_i|\mathbf{N}_i)} \tag{22}$$

for $i = 1, 2$.

We pointed out in Section 3 that majorities of experimental participants exhibit moderate risk aversion and rank-dependent utility. Because part of the point of the simulations is to model our theory for laboratory application, we specify $\hat{\omega}_{i|jklmn}$ in such a way as to allow for agents whose choices either respect Expected Utility Theory (EUT), or violate EUT axioms only in ways consistent with Rank-Dependent Utility Theory (RDU), as per Quiggin (1982). (RDU formally nests EUT.) A specification of RDU that has proven particularly useful in estimation of data from risky choice experiments is due to Prelec (1998), which is conventionally used to map a set of probability mass functions into a set of probability weighting functions to account for risk aversion. In the conditional game context, we use its most general (2-parameter) form

$$P(\omega) = \exp\big[-\eta(-\ln\omega)^{\varphi}\big] \;\; \varphi > 0, \eta > 0. \tag{23}$$

We assume that all agents are risk averse across all payoff intervals (though to degrees that can vary across the agents) and, therefore, use a strictly concave Prelec operator to transform $\hat{\omega}_{i|jklmn}$ into a utility weighting function

$$\bar{\omega}_{i|jklmn} = P(\hat{\omega}_{i|jklmn}) = \exp\big[-\eta(-\ln\hat{\omega}_{i|jklmn})^{\varphi}\big]. \tag{24}$$

The result is then normalized to define the conditional utility functions

$$\tilde{\omega}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_{jklmn}) = \frac{\bar{\omega}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_{jklmn})}{\bar{\omega}_{i|jklmn}(\mathsf{N}_i|\mathbf{N}_{jklmn}) + \bar{\omega}_{i|jklmn}(\neg\mathsf{N}_i|\mathbf{N}_{jklmn})} \tag{25}$$

for $i = 1, 2$.

In the context of unconditional utility, subjective decision weights are understood as reflecting idiosyncratic *beliefs* about probabilities of outcomes based on their utility ranking. In the praxeological context modeled by CGT, it is most natural to interpret the weighting function as reflecting the idea that an agent might strategically adjust the preferences expressed by actual *or* conjectured choices to reflect uncertainty about the extent to which those with whom she interacts are guided by the norm she anticipates. Since CGT utilities comply with the probability syntax, we may define a utility weighting function as analogous to a probability weighting function by transforming conditional utilities as developed above to account for this praxeological uncertainty.

In the simulations to follow we will *not* exploit the full flexibility of utility representation offered by the 2-parameter Prelec function. This is because the point of the simulations is to demonstrate the capacity of conditionalization to serve as a mechanism for effecting the kinds of normative dynamics Kuran identifies. For this purpose it is preferable to minimize other sources of complexity, so, as stated, we restrict attention to utility functions that are concave throughout the interval space. For now, we simply make the point that the theory can accommodate the more complicated functions (e.g., S-shaped, inverse S, and others; see Wilcox 2015) that often provide best fits to laboratory choice data. In empirical applications, identification of the Kuran parameters would require the experimenter to empirically estimate risk preferences and subjective probability weightings. As argued in Section 3, this is something experimenters interested in norms are motivated to do anyway.

5.2 Conditional Game Simulations of Normative Behavior

The Investment / Trust game in the six-agent network introduced in Section 4.2 provides a context for testing CGT modeling of relative norm fragility with Kuran utilities using computer simulations. The network graph (5) comprises two subnetworks, each of which begins (i.e. prior to conditionalisation) with a different governing norm in Bicchieri's sense. The subnetwork $\{X_1, X_2, X_3\}$ abides by norm N_1, Equality, and the subnetwork $\{Y_1, Y_2, Y_3\}$ abides by N_2, Equity. In the simulations, each agent chooses strategies that determine an action at each node of the extensive form of the game, on the assumption that they are equally

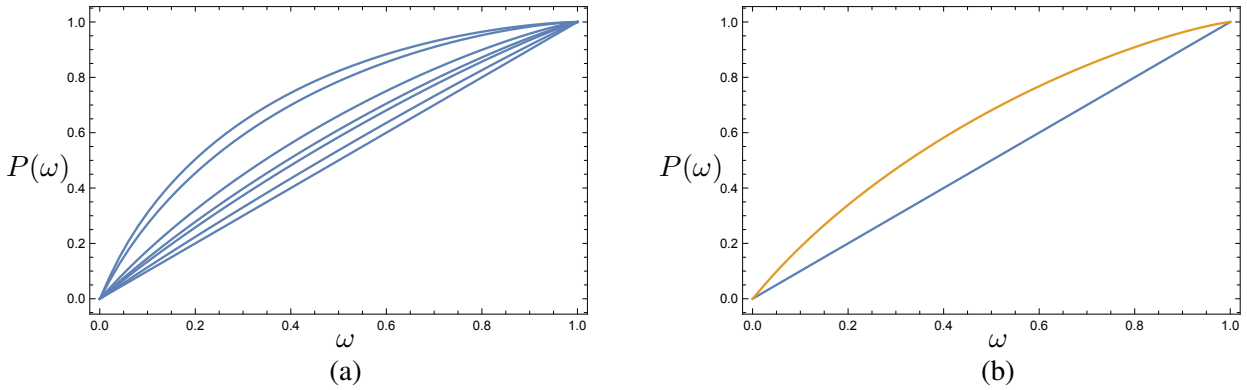


Figure 3: Prelec mapping functions: (a) random Prelec functions, (b) average Prelec function.

likely to find themselves in the Investor role or the Trustee role. The endowment for each agent in the Investor role is specified as $\mathcal{E} = \$12$ and the fraction of the endowment that may be offered is capped at $\sigma = 2/3$. Under both the Equality and Equity norms it is a dominant strategy for the Investor to choose the maximum possible transfer. This generates the following monetary payoffs:

$$\begin{aligned} I_{\mathsf{N}_1} &= (1 + 2\sigma)\mathcal{E}/2 = \$14 \\ T_{\mathsf{N}_1} &= (1 + 2\sigma)\mathcal{E}/2 = \$14\,, \end{aligned} \tag{26}$$

and

$$\begin{aligned} I_{\mathsf{N}_2} &= (1 - \sigma + 3\sigma^2)\mathcal{E} = \$20 \\ T_{\mathsf{N}_2} &= 3(\sigma - \sigma^2)\mathcal{E} = \$8\,. \end{aligned} \tag{27}$$

We assume, as stated in Section 5.2, that the agents are risk averse, which corresponds to a strictly concave Prelec function that crosses the diagonal at $\omega = 0$.[5] This crossing point is established for any given value of ϕ by solving for η according to the function

$$\eta = \exp[\log(-\ln \omega_c)(1 - \phi)]\,. \tag{28}$$

Our simulations are run with each agent assigned a Prelec function by randomly drawing ϕ from a uniform distribution over the interval $(1, 1.5)$. Figure 3(a) displays the realizations from the random draw of Prelec functions and Figure 3(b) displays the average Prelec function with parameters $(\bar{\phi}, \bar{\eta}) = (1.228, 0.6628)$. A representative agent's baseline utility is computed by considering intrinsic utility only, that is, with $\alpha = 10$ and $\beta = \delta = 0$, computed for the average Prelec parameters, yielding $\mu(I_{\mathsf{N}_1}) = 0.525$ and $\mu(I_{\mathsf{N}_2}) = 0.475$. Note that although expected monetary payoffs are the same for all players and under both norms, Equity involves higher variance, so, given the risk aversion built into $\tilde{\omega}$, $\mu(I_{\mathsf{N}_1)}) > \mu(I_{\mathsf{N}_2})$. (The representative agent is described only for expository purposes, playing no role in any simulations.)

We now demonstrate how application of CGT generates normative change for players with Kuran utility functions. We simulate three social scenarios. In each case, we depict two 3-agent 'communities' distinguished from one another by prevalence in each of an alternative norm regulating play in the Investment / Trust game. X agents follow the Equality norm and Y agents follow the Equity norm. We simulate encounters between the two communities, and apply conditionalization as the engine of their normative adjustments to one another.

[5] To avoid singularities, the zero crossing is set at $\omega_c = 0.0001$.

Harmony is a baseline scenario for comparison with subsequent more interesting ones. Here all agents are content with their respective normative status quo positions, in the sense that they earn δ expressive utility when they play according to their community's preferred norm.

Pluralistic Ignorance is a scenario in which Y agents are satisfied with their community's norm, but X agents are preference falsifiers, privately dispreferring their community's norm, but each unaware that their dissenting attitude is shared by their compatriots. Thus X agents trade off expressive utility for reputational utility, and their choices will be sensitive to the relative magnitudes of β and δ.

Activist/Trendsetter names a scenario that replicates **Harmony** except that one agent, Y_3, is an Activist (Kuran) or Trendsetter (Bicchieri) who does not have reputational utility as an argument in her utility function, but always multiplies expressive utility by 2δ. This agent will thus play against the norm preferred by both her native community and herself only when δ is below some threshold in relation to *I*.

We summarize the relationship between theory and our simulation setup. First, in every simulation players conditionalize on intrinsic utility, meaning here that their risk preferences influence one another. Specifically, Equity-governed agents will be conditioned, on encountering Equality-governed agents, to increase their probability of playing according to the Equality norm because they expect their partner to favour lower variance, and will accordingly attach greater weight to this preference in their own play. Players do not conditionalize on expressive utility, except in the special case we construct for the third simulation of a network that includes an Activist / Trendsetter. The basic mechanism of normative influence is conditionalization on reputational utility.

5.3 Harmony

We simulate four subscenarios, with results tabulated in Table 2. Because all players within each community have identical utility functions and expectations, the table shows outcomes for a representative agent from each community. The symmetry of the Harmony scenario facilitates intuitive introduction of a further degree of modeling freedom allowed by CGT, the extent to which 'visitors' to a normative community other than their own adjust their normative expectations. The scope in CGT to conditionalize, or not, on the reputational component of the utility function, and in one direction of influence or both, allows for representing the possibility that when in Rome I might not only do what the Romans do, but *approve* of Romans doing what the Romans do when they are in Rome (while I might disapprove of Romans doing what Romans do when they visit my community). We consider, then, two variants of the Harmony scenario:

Sovereign Communities: Agents award reputational utility only to choices expected under the norm of their home community: $R_{X_i}(N_i|N_j) = 0$ and $R_{Y_i}(N_i|N_j) = 0$ in (16) for all i, j.

Cosmopolitan Communities: Agents award reputational utility when agents play as expected according to the norm that governs their community by members of a community governed by a different norm: $R_{X_i}(N_i|N_j) = 1$ and $R_{Y_i}(N_i|N_j) = 1$ in (16) for all i, j.

In each subscenario, we simulate an instance of the two possible general inequalities between β and δ:

$\boldsymbol{\beta > \delta}$**:** Reputational utility dominates expressive utility ($\beta = 40$, $\delta = 4$).

$\boldsymbol{\beta < \delta}$**:** Expressive utility dominates reputational utility ($\beta = 4$, $\delta = 40$).

We show the results of the above simulations in Table 1. These can be summarized as follows. First, comparison of Table 1 with the baseline numbers shows the obvious result that adding reputational utility

and expressive utility as components of players' total utility increases the value to players in each community of playing according to their respective norms. More interestingly, when players conditionalize on reputational utility (that is, in the Cosmopolitan Communities subscenario) and the weight of reputational utility dominates the weight of expressive utility, followers of the less intrinsically valuable norm, Equity, earn higher expected utility from playing according to that norm when they visit members of the other community than they do by adapting their behavior to the local norm. Players whose preferred Equality norm earns higher expected intrinsic utility, on the other hand, prefer to trade off some of that gain in exchange for improved prospects when they meet Equity-governed counterparts. Mutual respect strengthens normative pluralism.

Table 1: Harmony simulation results.

	Sovereign Communities		Cosmopolitan Communities	
	$\beta = 40$	$\beta = 4$	$\beta = 40$	$\beta = 4$
Utility	$\delta = 4$	$\delta = 40$	$\delta = 4$	$\delta = 40$
$\overline{w}_X(\mathsf{N}_1)$	0.618	0.663	0.552	0.629
$\overline{w}_X(\mathsf{N}_2)$	0.382	0.337	0.448	0.371
$\overline{w}_Y(\mathsf{N}_1)$	0.427	0.381	0.516	0.410
$\overline{w}_Y(\mathsf{N}_2)$	0.573	0.619	0.484	0.590

5.4 Pluralistic Ignorance

We simulate a pluralistic ignorance scenario by assuming that X agents, privately disliking their community's Equality norm, can gain expressive utility only when they play according to the other community's norm. We investigate two subscenarios:

$\beta > \delta$: Reputational utility dominates expressive utility ($\beta = 40$, $\delta = 4$).

$\beta < \delta$: Expressive utility dominates reputational utility ($\beta = 4$, $\delta = 40$).

We simulate these subscenarios only for the 'Cosmopolitan Communities' environment, because it is trivial that in a 'Sovereign Communities' environment, preference falsifiers would simply swap reputational for expressive utility when they 'go abroad', so conditionalization would have no effect; any observed changes relative to the relevant comparison with the Harmony scenario would be entirely attributable to the interaction of risk aversion with the exogenously chosen β:δ ratio.

Table 2 shows the results for each subscenario. The key result is that, consistently with the prediction of the theory, when there is as much utility to be gained through expressive utility as through reputational utility, conditionalization results in a clearly observable shift toward play that follows the Equity norm, notwithstanding its higher risk. Our simulations under random assignments of Prelec weights as in the Harmony cases reveals, for each ratio of β:δ, specific thresholds at which preference falsifiers begin to behaviorally express their true preferences. For example, when β = δ, abandonment of preference falsification occurs between $\beta = \delta = 17$ and $\beta = \delta = 18$; when $\beta > \delta$, the threshold occurs at $\beta = 40$, $\delta = 5$. Thus, given known Prelec weights, β:δ ratios can be identified from observed behavior.

Analysis of this case confirms that modeling captures the theoretical target of interest, and shows that norms that are privately unpopular are relatively fragile in encounters with norms that are privately supported, even when, as here, there is no mechanism by which players can update their priors from interactions in their home network and learn directly about their pluralistic ignorance.

Table 2: Pluralistic Ignorance simulation results.

Utility	$\beta = 40$ $\delta = 4$	$\beta = 4$ $\delta = 40$
$\overline{w}_X(\mathsf{N}_1)$	0.504	0.409
$\overline{w}_X(\mathsf{N}_2)$	0.496	0.591
$\overline{w}_Y(\mathsf{N}_1)$	0.465	0.405
$\overline{w}_Y(\mathsf{N}_2)$	0.535	0.595

5.5 Activist/Trendsetter

Both Bicchieri (2017) and Kuran (1995) express interest in modeling the impact on normative dynamics of agents who are unconcerned with reputational utility but derive utility directly from converting others to their normative preferences. Bicchieri refers to such agents as 'trendsetters', and Kuran calls them 'activists'. We introduce an Activist / Trendsetter agent into the simulation environment by revising the basic utility function of one agent, Y_3, who possesses an idiosyncratic utility function, adopted from (16) for purposes, as follows:

$$\omega(I, E) = \omega(\alpha + I + \gamma E)\,. \tag{29}$$

For convenience in the simulation we arbitrarily set $\gamma = 2\delta$. Y_3 supports the norm of her group, but consistently with the concept of activism / trendsetting is unconcerned with reputational utility. This alteration by itself would not allow the Activist / Trendsetter to influence the utility functions of other agents because, in the austere informational conditions of the model, her pattern of play is not distinguishable from that of other Equity-governed agents. However, another aspect of activism as discussed by Kuran is that the activist derives utility directly from converting agents who do not share her norm to adoption of her normative point of view. We represent this in the model by means of the following device. We allow the activist agent to exert her 'missionary' influence by assigning δx units of expressive utility to Equality-governed agents (X group agents) when they play consistently with Y_3's favoured norm against Y_3 (but not when they play against non-activist Y-group agents). Thus, on this model, the presence of an activist induces a strictly limited element of ambiguity into the dispositions of agents she proselytises. This would make no coherent sense outside the context of conditionalization. But in the CGT framework, the ambiguity has a natural interpretation: the missionary activities of the activist cause the targets of her persuasion to conjecture states of the world in which they are persuaded by her, and in choosing actions to condition their preferences on those conjectures as well as on the conjecture that they remain loyal to their initial norm. This model thus nicely illustrates an aspect of modeling power that is special to CGT.

In all other respects, agents in this scenario are assigned the same utility structure as in the Harmony scenario.

By placing the activist / trendsetter in the community governed by a norm that mandates the behavior that earns riskier intrinsic utility, we ensure that any behavioral change we observe against the Harmony scenario must be driven by conditionalization on the non-monetary payoffs. To investigate this possible effect we simulate the same environments (Cosmopolitan and Sovereign communities) as in the Harmony simulations, with the exception that the Activist agent Y_3 does not condition on reputational utility in the Cosmopolitan communities scenario because this is not an argument in her utility function. In each environment, we simulate an instance of the two possible general inequalities between β and δ:

$\boldsymbol{\beta > \delta}$**:** Reputational utility dominates expressive utility ($\beta = 40$, $\delta = 4$).

$\boldsymbol{\beta < \delta}$**:** Expressive utility dominates reputational utility ($\beta = 4$, $\delta = 40$).

Table 3 displays the simulation results. Since the X community is homogenous, we show results for a representative X agent. To interpret this table we must compare it to the results of the Harmony case as displayed in Table 1. Upon comparing the utility values for the X agents, it is clear that under all conditions, the X-group agents in the Activist / Trendsetter case place more weight on play governed by the equity norm. However, because the margins are very small, the changes cannot be regarded as economically significant. To cope with this problem, we simulate a case that does *not* correspond to the activism phenomenon as understood by Kuran, which involves an activist *minority*. We simulate scenarios in which activists are a *majority* in the *Y*-group, i.e., there are two activists instead of one. The point of this model is to license an inference. Here we observe significant loss of support for the Equality norm among *X*-group players in both the Cosmopolitan communities and Sovereign communities environments and for both inequalities between the Kuran parameters. Since nothing but conditionalization can be driving these changes, and the only difference between the 1-activist simulation and the 2-activist simulation is in the number of activists, we can deduce that the activist mechanism is generating the theoretically expected effect.

These scenarios serve as demonstrations of the capacity of CGT to represent and facilitate estimation of mechanisms of normative influence and diffusion. We have shown how three of Kuran's instances of normative change at the social level could be identified in possible choice data. The mechanism used to achieve this identification is conditionalization of preference mediated by both descriptive and normative expectations. Thus, we contend, we have operationalized Bicchieri's analysis of norms for potential use by empirical political scientists or economists presented with controlled observations of choice behavior.

Table 3: Activist/Trendsetter simulation results.

	One activist				Two activists			
	Cosmopolitan community		Sovereign community		Cosmopolitan community		Sovereign community	
Utility	$\beta = 40$ $\delta = 4$	$\beta = 4$ $\delta = 40$	$\beta = 40$ $\delta = 4$	$\beta = 4$ $\delta = 40$	$\beta = 40$ $\delta = 4$	$\beta = 4$ $\delta = 40$	$\beta = 40$ $\delta = 4$	$\beta = 4$ $\delta = 40$
$\overline{w}_X(\mathsf{N}_1)$	0.545	0.611	0.545	0.611	0.543	0.592	0.543	0.592
$\overline{w}_X(\mathsf{N}_2)$	0.455	0.389	0.455	0.389	0.457	0.408	0.457	0.408
$\overline{w}_{Y_1}(\mathsf{N}_1)$	0.513	0.410	0.423	0.400	0.513	0.410	0.423	0.400
$\overline{w}_{Y_1}(\mathsf{N}_2)$	0.487	0.590	0.577	0.600	0.487	0.590	0.577	0.600
$\overline{w}_{Y_2}(\mathsf{N}_1)$	0.513	0.410	0.423	0.400	0.509	0.404	0.509	0.405
$\overline{w}_{Y_2}(\mathsf{N}_2)$	0.487	0.590	0.577	0.600	0.491	0.596	0.491	0.596
$\overline{w}_{Y_3}(\mathsf{N}_1)$	0.491	0.355	0.491	0.355	0.391	0.35	0.491	0.355
$\overline{w}_{Y_3}(\mathsf{N}_2)$	0.509	0.645	0.509	0.645	0.50	0.645	0.509	0.645

6 Conclusion

A fully general theory of norms that can be applied to empirical data, and in particular to data generated by choice experiments, remains an outstanding goal that must consolidate the following contributions:

1. a satisfactory philosophical analysis of the concept of a norm;

2. a relatively general economic theory that links norms as social structures with incentives that motivate normatively regulated agents' choices in small-*n* scenarios where agents influence what Kuran calls 'intrinsic utility' and cannot be modeled as norm-takers;

3. a menu of standard experimental and econometric estimation procedures that are aligned with (1) and (2).

In this chapter, we have not focused on goal (1). Bicchieri's philosophical analysis may be overdemanding in requiring *fully* aligned expectations of descriptive and normative beliefs, but it is not clear how this criterion might best be relaxed without removing the teeth from its bite. However, we have proceeded on the assumption that Bicchieri's basic insight that norms are networks of expectations is correct, provided it is consistently given a genuinely social rather than an implicitly psychological interpretation as we urge in Section 2.

We believe that Kuran has provided a useful high-level economic model of norms for large-*n* scenarios in which agents are norm-takers and cannot influence their own utility except through their choice between falsifying and not falsifying their preferences, or adopting activist attitudes. This model is obviously not general in failing to apply to small-*n* interactions. In addition it is defined only over low-information representations of utility functions as categorical preference orderings. In the chapter we have taken a step toward greater generality on the second dimension by incorporating risk attitudes and subjective probability weighting into the basic Kuran utility model. We did this not for its own sake but because of our primary interest in adopting Kuran's theory to the kinds of small-*n* interactions that occur in the experimental laboratory.

This goal reflects our aim to have, at the very least, shown social scientists some of what will be required theoretically and conceptually if norms are to become direct objects of experimental study in themselves. It is particularly political scientists who are the intended audience here, as many sociologists might view norms, even if not conceived as fundamentally psychological, as too grounded in intentional structure and attitudes to be good constructs for integration into their theoretical space. Perhaps in that case, however, our modeling apparatus of CGT may carry some appeal. Though CGT, like any extension of game theory, operates on utility, agency, and choice, it represents these concepts as fundamentally social, indeed as, effectively, distributions in populations of dispositions to be influenced in certain ways.

It seems difficult, from any disciplinary perspective, to imagine fully modeling political dynamics as independent of normative identities, normative commitments of varying degrees of flexibility, and energies marshalled for the exercise of normative influence. We hope in this chapter to have offered some tools for such modeling that political scientists will want to refine, adapt, and ultimately apply to data.

Summarizing the tools in question, probability theory is an ideal mechanism with which to model dynamics that are fundamentally driven by weights of relative influence. Merging probability theory with network theory through the structure of Bayesian networks serves as natural syntax for representing a mechanism by which normative influence diffuses throughout a community. Standard Bayesian network theory is restricted to acyclic networks, but CGT relaxes that restriction by incorporating recognition that a cycle can be modeled as an infinite time sequence of acyclic networks. Thus, a cyclic network can be modeled as a Markov chain, and the Markov chain convergence theorem establishes necessary and sufficient conditions for convergence. The usefulness of this theorem is further enhanced by the fact that the converged state can be derived from the closed-loop transition matrix without having to conduct or trace iterations in literal time. We do not deny that people generally learn about norms and their effects by encountering one another in sequences of interactions in real time, and updating their expectations on the basis of such experience. But it is frequently inconvenient or impossible to experimentally set up such dynamics for controlled observation. In such circumstances the experimenter may need a representation of dynamics in the limit to compare with those she observes in her lab. We will have succeeded in our main aim if, when that need arises, she finds value in the resources we have provided.

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 path is a* cycle *if* $X_j \mapsto X_j$*. 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* $\boldsymbol{\mathcal{U}} := \{\tilde{u}_{i|\mathrm{pa}(i)},\ i = 1, \ldots, n\}$ *is the set of* conditional utilities *such that* $\tilde{u}_{i|\mathrm{pa}(i)}$ *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* $\tilde{u}_{i|\mathrm{pa}(i)}(\cdot|\boldsymbol{\alpha}_{\mathrm{pa}(i)})$*, is an ordering function such that, given the antecedent* $\mathrm{pa}\,(X_i) \vDash \boldsymbol{\alpha}_{\mathrm{pa}(i)}$*, then*

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

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* $\tilde{u}_{i|\mathrm{pa}(i)}(a_i|\boldsymbol{\alpha}_{\mathrm{pa}(i)}) = \tilde{u}_i(a_i)$*, 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,

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

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

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

where $(a_1, \ldots, a_n)$, termed the *coordination profile*, is the set of self-conjectures of $\{X_1, \ldots, X_n\}$. If $\tilde{u}_{i|\mathrm{pa}(i)}(a_i) = \tilde{u}_i(a_i)$, 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 $\tilde{u}_{i|\mathrm{pa}(i)}$ 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

$$X_1 \underset{\tilde{u}_{1|2}}{\overset{\tilde{u}_{2|1}}{\rightleftarrows}} X_2 \tag{A.6}$$

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 $\{Y(s), s \in \{1, 2, \ldots\}$ of the form

$$Y(1) \xrightarrow[p_{2|1,\, s=1}]{} Y(2) \xrightarrow[p_{3|2,\, s=2}]{} Y(3) \xrightarrow[p_{4|3,\, s=3}]{} Y(4) \xrightarrow[p_{5|4\,, s=4}]{} \cdots \,, \tag{A.7}$$

where $p_{s+1|s}$ is the conditional probability mass function governing Y_{s+1} given Y_s. This probability structure assures that Y_{s-1} and $Y s +_1$ are conditionally independent, given Y_s. In other words, the Markov property

is equivalent to the statement that the state of 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

$$X_1(1) \xrightarrow[\tilde{u}_{2|1,\ s=1}]{} X_2(2) \xrightarrow[\tilde{u}_{1|2,\ s=2}]{} X_1(3) \xrightarrow[\tilde{u}_{2|1,\ s=3}]{} X_2(4) \xrightarrow[\tilde{u}_{1|2\ ,s=4}]{} \cdots . \tag{A.8}$$

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

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

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

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

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

with marginal for X_2 computed at time $s = 2$ 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

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

where the *mass vector* is

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

and

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

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 time s is

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

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

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

where t is expressed in closed-loop time 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*

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

and c)

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

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).

A.4 Six-Agent Cyclic Network Markov Equivalent Derivation

Our task in this section is to establish conditions such that (10) is Markov equivalent with (4) and to define the transition matrices $T_{i|jklmn}$ and $T_{ijklm|jklmn}$ where the subscripts are members of the index sets (5) and (11). Following the definition of $T_{i|j}$ in (A.14), transition matrix $T_{i|jklmn}$ is the 2×32 matrix

$$T_{i|jklmn} = \begin{bmatrix} \tilde{u}_{i|jklmn}(x_{i1}|x_{j1}, x_{k1}, x_{l1}, x_{n1}, x_{n1}) & \cdots & \tilde{u}_{i|jklmn}(x_{i1}|x_{j2}, x_{k2}, x_{l2}, x_{n2}, x_{n2}) \\ \tilde{u}_{i|jklmn}(x_{i2}|x_{j1}, x_{k1}, x_{l1}, x_{n1}, x_{n1}) & \cdots & \tilde{u}_{i|jklmn}(x_{i1}|x_{j2}, x_{k2}, x_{l2}, x_{n2}, x_{n2}) \end{bmatrix} \tag{A.19}$$

The entries of $T_{ijklm|jklmn}$ are conditional coordination functions of the form $w_{ijklm|jklmn}$ for the subnetwork $\{X_i, X_j, X_k, X_l, X_m\}$ given the subnetwork $\{X_j, X_k, X_l, X_m, X_n\}$. Suppressing arguments and applying the chain rule, we obtain

$$w_{ijklm|jklmn} = w_{m|ijkljklmn} w_{ijkl|jklmn} \tag{A.20}$$

The conditional mass function $w_{m|ijkljklmn}$, however, involves self-conditioning for X_m, and thus is a degenerate mass function of the form (eliminating redundant conditioning indices)

$$w_{m|ijklmn}(a_m|a_i, a_j, a_k, a_l, a'_m, a_n) = \begin{cases} 1 & \text{if } a_m = a'_m \\ 0 & \text{otherwise.} \end{cases} \tag{A.21}$$

Applying the chain rule to $w_{ijkl|jklmn}$ yields

$$w_{ijkl|jklmn} = w_{l|ijkjklmn} w_{jkl|jklmn}, \tag{A.22}$$

where $w_{l|ijkjklmn}$ involves self-conditioning for X_l hence

$$w_{l|ijklmn}(a_l|a_i, a_j, a_k, a'_l, a_m, a_n) = \begin{cases} 1 & \text{if } a_l = a'_l \\ 0 & \text{otherwise.} \end{cases} \tag{A.23}$$

Continuing this process, it follows that

$$w_{k|ijklmn}(a_k|a_i, a_j, a'_k, a_l, a_m, a_n) = \begin{cases} 1 & \text{if } a_k = a'_k \\ 0 & \text{otherwise,} \end{cases} \tag{A.24}$$

and

$$w_{j|ijklmn}(a_j|a_i, a'_j, a_k, a_l, a_m, a_n) = \begin{cases} 1 & \text{if } a_j = a'_j \\ 0 & \text{otherwise.} \end{cases} \tag{A.25}$$

Combining all terms,

$$w_{ijkl|jklmn}(a_i, a_j, a_k, a_l, a_m | a'_j, a'_k, , a'_l, a'_m, a_n) = \begin{cases} w_{i|jklmn}(a_i | a_j, a_k, a_l, a_m, a_n) & \text{if } a_j = a'_j, a_k = a'_k, a_l = a'_l, a_m = a'_m \\ 0 & \text{otherwise.} \end{cases} \tag{A.26}$$

These valuations are used to populate the subnetwork-to-subnetwork transition matrices from $\{X_j, X_k, X_l, X_m, X_n\}$ to $\{X_i, X_j, X_k, X_l, X_m\}$, thereby completing the cycle defined by 10).

References

Akerlof, G. & Kranton, R. (2010). *Identity Economics*. Princeton University Press.

Andersen, S., Harrison, G., Lau, M., & Rutström, E.E. (2008). Lost in state space: Are preferences stable? *International Economic Review* 49: 1091-1112.

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

Andreoni, J., Nikiforakis, N. & Siegenthaler, S. (2017). Social change and the conformity trap. Working paper, *Semantic Scholar*: https://www.semanticscholar.org/paper/Social-Change-and-the-Conformity-Trap-

Andrighetto, G., Grieco, D., & Tummolini, L. (2015). Perceived legitimacy of normative expectations motivates compliance with social norms when nobody is watching. *Frontiers in Psychology* 6: 1413.

Aumann, R.J. (1987). Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society*, 1-18.

Azar, O.H. (2018). The influence of psychological game theory. *Journal of Economic Behavior & Organization* 167: 445-453.

Banerjee, A. (1992). A simple model of herd behavior. *Quarterly Journal of Economics* 107: 797-817.

Battigalli, P. & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2): 170-176.

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

Berg, J., Dickhaut, J. & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior* 10: 122-142.

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

Bicchieri, C. (1993). *Rationality and Coordination*. Cambridge University Press.

Bicchieri, C. (2006). *The Grammar of Society–The Nature and Dynamics of Social Norms*. Cambridge University Press.

Bicchieri, C. (2017). *Norms in the Wild*. Cambridge University Press.

Bicchieri, C. & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making* 23: 161-178.

Bicchieri, C. & Chavez, A. (2013). Norm manipulation, norm evasion: Experimental evidence. *Economics and Philosophy* 29: 175-198.

Bicchieri, C., Lindemans, J., & Jiang, T. (2014). A structured approach to a diagnostic of collective practices. *Frontiers of Psychology* 5: 1418 doi: 10.3389/fpsyg.2014.01418

Bicchieri, C., & Sontuoso, A. (2017). Game-theoretic accounts of social norms: The role of normative expectations. PPE Working Papers 0011, Philosophy, Politics and Economics, University of Pennsylvania. https://ideas.repec.org/s/ppc/wpaper.html

Bicchieri, C. & Xiao, E. (2009). Do the right thing: But only if others do so. *Journal of Behavioral Decision Making* 22: 191-208.

Bicchieri, C., Xiao, E. & Muldoon, R. (2011). Trustworthiness is a social norm, but trusting is not. *Politics, Philosophy and Economics* 10: 170-187.

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

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

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

Binmore, K. (2007). *Does Game Theory Work? The Bargaining Challenge.* MIT Press.

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

Brock, W. & Durlauf, S. (2001). Discrete choices with social interactions. *Review of Economic Studies* 68: 235-260.

Castelfranchi, C. (2010). *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley.

Chambers, R., & Quiggin, J. (2000). *Uncertainty, Production, Choice, and Agency: The State-Contingent Approach.* Cambridge University Press.

Chamley, C. (2004). *Rational Herds*. Cambridge University Press.

Chetty, R., Hofmeyr, A., Kincaid, H. & Monroe, B. (2020). The Trust game does not (only) measure trust: The risk-trust confound revisited. *Journal of Behavioral and Experimental Economics*, forthcoming. DOI: 10.1016/j.socec.2020.101520

Cowell, R.G., Dawid, A.P. , Lauritzen, S.L. & Spiegelhalter, D.J. (1999). *Probabilistic Networks and Expert Systems*. Springer Verlag.

Druckman, J., Green, D., Kuklinski, J. & Lupia, A., eds., (2011). *Cambridge Handbook of Experimental Political Science.* Cambridge University Press.

Duffy, J. & Lafky, J. (2019). Social conformity under evolving private preferences. Working paper: https://www.socsci.uci.edu/ duffy/papers/LivingLie12132019.pdf

Ensminger, J. & Henrich, J., eds. (2014). *Experimenting With Social Norms*. Russell Sage Foundation.

Gerber, A.S. & Rogers, T. (2009). Descriptive social norms and motivation to vote: Everybody's voting and so should you. *Journal of Politics* 71: 178-191.

Gintis, H. (2014). *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences-Revised Edition*. Princeton University Press.

Guala, F. (2016). *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton University Press.

Hanneford, I. (1995). *Race: The History of an Idea in the West.* Johns Hopkins University Press.

Harrison, G. (2014). Real choices and hypothetical choices. In S. Hess & A. Daly, eds., *Handbook of Choice Modelling*, pp. 236-254. Edward Elgar.

Harrison, G., Martínez-Correa, J., Swarthout, J.T. & Ulm, E. (2017). Scoring rules for subjective probability distributions. *Journal of Economic Behavior & Organization* 134: 430–448.

Harrison, G. & Ross, D. (2016). The psychology of human risk preferences and vulnerability to scare-mongers: Experimental economic tools for hypothesis formulation and testing. *Journal of Cognition and Culture* 16: 383-414.

Harrison, G. & Rutström, E. (2008). Risk aversion in the laboratory. In J. Cox & G. Harrison, eds., *Risk Aversion in Experiments*, pp. 41–197. Emerald.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H., eds. (2004). *Foundations of Human Sociality*. Oxford University Press.

Hirscheifer, J., & Riley, J. (1992). *The Analytics of Uncertainty and Information*. Cambridge University Press.

Kagel, J. & Roth, A. (2016). *The Handbook of Experimental Economics, Volume 2*. Princeton University Press.

Karni, E. (1990). State-dependent preferences. In J. Eatwell, M. Milgate, & P. Newman, eds. *The New Palgrave: Utility and Probability*, pp. 242-247. Norton.

Katz, D. & Allport, F. (1931). *Student Attitudes*. Craftsman.

Kreps, D. (1990). *Game Theory and Economic Modelling*. Oxford University Press.

Kuran, T. (1995). *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press.

Lerner, B. (2011). *One For the Road: Drunk Driving Since 1900.* Johns Hopkins University Press.

Lewis, D. (1969). *Convention: A Philosophical Study*. Harvard University Press.

Lewis, D. (1976). Convention: Reply to Jamieson. *Canadian Journal of Philosophy*, 6(1):113–120.

Lewis, M. (2018). *The Fifth Risk.* Norton.

Luenberger, D.G. (1979). *Introduction to Dynamic Systems*. John Wiley.

Mansbridge, J., Hartz-Karp, J., Amengual, M. & Gastil, J.(2006). Norms of deliberation: An inductive study. *Journal of Public Deliberation* 2:7.

Manski, C.F. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives*, 14 : 115–136.

Martin, J.L. (2009). *Social Structures*. Princeton University Press.

Matheson, J.E. & Winkler, R.L. (1976). Scoring rules for continuous probability distributions. *Management Science* 22: 1087–1096.

Michaeli, M. & Spiro, D. (2015). Norm conformity across societies. *Journal of Public Economics* 132: 51-65.

Michaeli, M. & Spiro, D. (2017). From peer pressure to biased norms. *American Economic Journal: Microeconomics* 9: 152-216.

Murray, J.A.H., Bradley, H., Craigie, W.A. & Onions, C.T., eds. (1991). *The Compact Oxford English Dictionary*. Oxford University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Prelec, D. (1998). The probability weighting function. *Econometrica*, 60:497–528.

Quiggin, J. (1982). A theory of anticipated utility. Journal of Economic Behavior and Organization 3: 323–343.

Rizzolatti, G., Forgassi, L. & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Neuroscience Reviews*, 2:661–670.

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

Ross, D. & Stirling, W.C. (2020). Economics, social neuroscience, and mindshaping. In J. Harbecke & C. Hermann-Pillath, eds., *The Brain and the Social–Methods and Philosophy of Integrating Neuroscience and Economics*, chapter 10. Routledge. in preparation.

Schelling, T.C. (1980). *The Strategy of Conflict*. Harvard University Press.

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

Stirling, W.C. (2016). *Theory of Social Choice on Networks*. Cambridge University Press.

Sugden, B. (1986/2004). *The Economics of Rights, Co-operation and Welfare*. Palgrave Macmillan.

Tummolini, L., Andrighetto, G., Castelfranchi, C. & Conte, R. (2013). A convention or (tacit) agreement betwixt us: On reliance and its normative consequences. *Synthese*, 190(4):585–618.

Wilcox, N. (2015). Unusual estimates of probability weighting functions. *ESI Working Paper 15-10*. Retrieved from http://digitalcommons.chapman.edu/esi-working-papers/159.

Wilson, R. & Eckel, C. (2011). Trust and social exchange. In J. Druckman, D, Green, J. Kuklanski, & A, Lupia, eds., *Cambridge Handbook of Experimental Political Science*, pp. 243-257. Cambridge University Press.

Xiao, E. & Bicchieri, C. (2010). When equality trumps reciprocity. *Journal of Economic Psychology* 31: 456-470.

Zawidzki, T.W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press.