

Mindshaping, Conditional Games, and the Harsanyi Doctrine

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

Brigham Young University

March 2, 2023

Abstract

Much of the game theory literature concerns mechanisms by which players can infer information about the utilities, beliefs, and strategies of other players based on actions within games and pre-play signals. When game theory is applied to strategic interactions among people, such analysis interprets them as trying to “mindread”. Recent work in cognitive science, however, suggests that human coordination rests more centrally and necessarily on “mindshaping” processes, in which people resolve equivocal preference content jointly. This kind of process cannot be modeled using standard resources of game theory. However, as mindshaping is strategic, there is motivation to widen the game theory toolkit to accommodate it. Conditional Game Theory is a strategic theory of mindshaping. We show how it can be used to help players of standard games identify correlated equilibrium, and thus solve games. We then extend CGT to address a challenge to the relevance of correlated equilibrium to empirical choice data. This is based on the fact that correlated equilibrium requires the Harsanyi Doctrine, according to which Bayesian players share common priors; but the majority of observed empirical choice behavior under risk violates this Doctrine. We show how pre-play analysis using CGT can reconcile the Harsanyi Doctrine with rank-dependent choice as typically seen in economic experiments.

1 Introduction

In both abstract and applied game theory, it is typically assumed that equilibrium selection requires imperfectly informed players to infer one another’s most probable patterns of play. The nature of such inferences varies with the specified information environment: players might know one another’s preference orderings but need to estimate risk attitudes, or they might be entirely ignorant about one another’s utility functions before observing some actions. In such cases the analyst

models a process that is analogous to what cognitive scientists call ‘mindreading’ (Nichols and Stich, 2002). A capacity for making inferences about others’ states of mind, particularly beliefs and preferences, is taken by many behavioral and social scientists to be a prerequisite for sophisticated social cooperation as well as for strategies based on deliberate deception. Special human mindreading dispositions and abilities are often claimed to have been crucial to the achievement of ecological dominance by *Homo sapiens* (Malle et al., 2003; Frith and Wolpert, 2004).

The relationship between natural mindreading and theoretical inferences of beliefs and preferences in game theory is merely one of analogy. The developers of game theory have not taken themselves to be building an empirically inspired theory of mindreading, and proponents of mindreading in cognitive and behavioral science have not been motivated by game theory. However, we suggest that confidence in natural mindreading capacity is important to motivating the plausibility of many applications of game theory to modeling small-scale interactions among people. In particular, the mindreading hypothesis makes the idea that people can jointly arrive at a selected equilibrium in multi-equilibrium games non-mysterious, even if it leaves the real work of technical specification of equilibrium selection principles still to be done by the game theorist.

Recently, however, Zawidzki (2013) and other cognitive scientists have argued persuasively that the importance of mindreading to human social development and coordination has been significantly overstated. Mindreading, these scientists argue, is often too severely underdetermined by available evidence to be feasible in cases where people nevertheless achieve relatively precise alignments of mutual expectations. The primary family of mechanisms for such coordination, Zawidzki argues, is *mindshaping*. This family is united by a basic disposition among people to influence one another to dynamically *form* shared complexes of belief and preference. Specific mindshaping mechanisms include imitation, deliberate pedagogy, specialization and exchange of knowledge, and leverage based on differentiated social status and roles. Such mindreading as people who know one another well sometimes achieve is argued by Zawidzki to be parasitic on a foundation of mindshaping.

In one key respect the general idea of mindshaping is not foreign to game theorists. Mindshaping mechanisms all involve public signaling. For example, millions come to share the belief that Ukraine is a victim because they all share in observing multiple public expressions, in print and speech, of this idea. This shared belief is then available to motivate expectations that proposals to aid Ukrainians will find coordinated responses. In game theoretic terms, the shared signal functions as a (partial) device for correlated equilibrium (Aumann, 1974, 1987). Under certain conditions, discussed below, correlated equilibrium is a powerful device for solving games.

In another respect, however, mindshaping is at odds with the intuitions underlying standard applications of game theory to social coordination. In games where players must infer preferences and beliefs of other agents, they know their *own* preferences and beliefs. Furthermore, they assume that the set of others’ preferences and beliefs they work to infer are fixed points (or, in cases of beliefs about probabilities, or beliefs structured as confidence intervals, or preferences over continuous scales, fixed distributions). But mindshaping is constructed as a dynamic that *changes* preferences and beliefs. Furthermore, and more radically, models of mindshaping depict agents who are *ex ante* unsure of their own preferences and beliefs until the process of mutual accommodation settles into an *ex post* equilibrium. Indeed, mindshaping theorists generally also favor philosophical *externalism* about beliefs and preferences (Dennett, 1987; Clark, 1997; Hutto, 2008), according to which these “propositional attitudes” are social constructions for interpreting behavior (including people’s interpretation of their own behavior), and have no exact counterparts

at the scale of ‘latent’ individual brain states. Such externalism is not a niche view among philosophers of cognitive science; for several decades it has been growing into the dominant position. According to externalism, the mindreading hypothesis is not just empirically exaggerated but conceptually confused, as it imagines that people generally engage in inferences about ‘private’ states of mind that exist only in cases where people think by explicitly attending to fragments of silently rehearsed English or Chinese or Swahili etc..

Externalism, and social explanation based on mindshaping rather than mindreading, need not *directly* trouble game theorists. Again, their enterprise is not mathematical social psychology. However, mindshaping and externalism undermine the idea that people selecting equilibria in strategic scenarios modeled as games can do so because they can infer latent preferences and beliefs. Furthermore, because mindshaping describes processes of preference *change*, such a process cannot be modeled as an extensive-form game soluble by backward induction or sequential equilibrium (Kreps and Wilson, 1982). These solution procedures *begin* from outcomes specified as vectors of utility indices that hold across all information sets in a game tree. Thus the outcome values to players cannot be generated *ex post* by the play of the game. Mindshaping shares this awkward aspect with the concept of preference construction in experimental psychology (Lichtenstein and Slovic, 2006) (though the ideas are not the same, and we will say no more about preference construction here), which has resisted integration with behavioral game theory (Camerer, 2003).

We aim to show that the tension just identified, far from being a conundrum or a barrier to the smooth transfer of intuitions between game theory and experimental social science, is a source of productive insights. The affinities between mindshaping and correlated equilibrium are more important than apparent discrepancies in concepts of mental states.

Specifically, our objective here is to show how *Conditional Game Theory* (CGT) (Stirling, 2012, 2016), which we interpret as a formal theory of mindshaping (Ross and Stirling, 2021), can be used to support the plausibility of the conditions for correlated equilibrium, at least in applications where individual people are players. This is based on new technical extensions to CGT that go beyond Stirling’s original formulation. Furthermore, one of these extensions, to take account of widespread empirical observations about the structure of human risk preferences, allows us to offer a new answer to an empirically based problem for the applicability of correlated equilibrium solutions to behavioral games.

The paper is structured as follows. In Section 2 we summarise the basic elements of CGT, including extensions beyond Stirling (2012) that are introduced in Ross and Stirling (2021). In Section 3 we argue that CGT analysis of players conditioning on pre-play actions can be used to generate shared signals about unconditional utilities, which can support assumption of common priors in Bayesian game players (the Harsanyi Doctrine) and allow them to identify correlated equilibrium as per the argument of Aumann (1987). Aumann’s argument depends on the assumption that all players conform to the expected utility axioms of Savage (1954). This might seem to make the argument uninteresting to experimenters, since more observed choice behavior coheres with Rank Dependent Utility (RDU) theory than with EUT. However, in Section 4 we show that CGT can incorporate RDU in the pre-play analysis, and reflect its effects in the transition matrices that game players can identify. In the concluding Section 5 we interpret this to suggest that Aumann’s argument can be applied to empirical choice data after all, at least where it can be supplemented by CGT analysis.

2 Conditional Game Theory

The basic structural condition for mindshaping is that agents are socially connected to each other such that each agent’s beliefs and preferences are influenced by the beliefs and preferences of others in the network. We refer to an interacting group of agents where this condition applies as a *social influence network*. We assume that interactions in a social influence network are necessarily strategic. This assumption may seem surprising. Could we not imagine a multi-agent network in which agents are programmed to generate and receive influence deterministically, without regard to their own prior utilities? Some of the ‘swarm intelligence’ literature in artificial intelligence (e.g. Bonabeau et al. (1999)) has this characteristic, which is often informally motivated by alleged facts about individually simple agents such as social insects. However, following Dennett (2017), there is no scientific payoff to ascribing beliefs and preferences to such fully programmed agents, and the need to make such ascriptions to explain behavior is necessary and sufficient for identifying a locus of mind. (The stylized fact about social insects as mindless robots is false; see Chittka (2022).) Therefore, mindshaping occurs only where agents are guided by interests and perspectives of their own. Consequently, mindshaping always involves a strategic aspect, and so in principle game theory should be relevant to analyzing it.

CGT was originally developed by Stirling (2012) for application to distributed control designs in AI. Such architectures achieve no efficiency gains if sub-systems have no autonomy at all. On the other hand, they are likely to fail design specifications if the sub-systems are not ultimately forced to align in choices of joint actions. This control restriction is absent in standard non-cooperative game theory, where equilibrium solutions can sometimes be catastrophically inefficient (as in tragedies of the commons, for example).

The challenge to which CGT responds is the need to represent agents as having strategic interests without specifying their ‘final’ utilities *ex ante*. This is necessary to allow for their preferences to be modified by social influence. CGT consequently models social influence networks by means of *conditional utilities* that are defined over *actions* rather than *outcomes* (Stirling, 2012, 2016). Of course, solutions still require reference to outcomes, so these cannot be neglected in the final analysis. As in standard noncooperative game theory, *categorical utilities* are defined over outcomes that are vectors of strategies, one for each agent in a game.

The foundational existence result for CGT establishes that the dynamic social interaction generated from conditional utilities as players participate in a conditional game converges to steady-state quantities termed *complementary coordination functions* that account for all social interrelationships that exist among each agent’s *complementary subgroup*, that is, the subset of all agents in a social influence network excluding the agent under consideration. These coordination functions can be used to generate unconditional individual utility functions defined over outcomes that send signals for social influence. We can crisply state the conceptual innovation this way: in a conditional game, signals don’t just communicate information about preferences, but change them. Conditional games are pre-play processes that *determine* subsequently analyzed standard games (Ross 2004, 2005, 2006).

2.1 Conditional Utilities

The key technical distinction between standard noncooperative game theory and conditional game theory is marked in the mathematical structure of utilities. The utilities of standard game theory are

categorical – each player’s utility function is fixed, unconditional, and defined over *outcomes*, that is, the joint actions of all players. Any social influence that the players exert on each other must be taken to be already incorporated into their utility functions. In the CGT setup, by contrast, players’ utilities are *conditional* and dependent on actions of others. As they engage in conditional play, their conditional utilities propagate throughout the network and generate emergent community level social relationships. In this section we summarize the basic mathematics of this modeling.

To set notation, we first consider standard noncooperative game theory and then extend definitions to the space of conditional games.

Definition 1 A noncooperative normal-form game is a triple $\{\mathbf{X}, \mathcal{A}, \mathcal{U}\}$ comprised of the following elements:

Agents: $\mathbf{X} = \{X_1, \dots, X_n\}$, a group of agents with each X_i possessing a finite action space $\mathcal{A}_i = \{x_{i1}, \dots, x_{iN_i}\}$, $i = 1, \dots, n$.

Outcome space: The Cartesian product $\mathcal{A} = \mathcal{A}_1 \times \dots \times \mathcal{A}_n$ is termed the outcome space. The elements of \mathcal{A} are enumerated as

$$\begin{aligned} \mathcal{A} = \{ \mathbf{x}_{k_1 k_2 \dots k_{(n-1)k_n}} = (x_{1k_1}, x_{2k_2}, \dots, x_{(n-1)k_{n-1}}, x_{nk_n}) \\ \in \mathcal{A}_1 \times \mathcal{A}_2 \times \dots \times \mathcal{A}_{n-1} \times \mathcal{A}_n, \\ k_i = 1, \dots, N_i, i = 1, \dots, n \}, \quad (1) \end{aligned}$$

where the indices k_1, \dots, k_n are incremented in lexicographical order from k_n to k_1 for $k_1 = 1, \dots, N_1$, $k_2 = 1, \dots, N_2$, through $k_n = 1, \dots, N_n$, yielding

$$\begin{aligned} \mathcal{A} = \left\{ \underbrace{(x_{11}, x_{21}, \dots, x_{(n-1)1}, x_{n1})}_{\mathbf{x}_{11\dots11}}, \underbrace{(x_{11}, x_{21}, \dots, x_{(n-1)1}, x_{n2})}_{\mathbf{x}_{11\dots12}}, \right. \\ \left. \dots, \underbrace{(x_{1N_1}, x_{2N_2}, \dots, x_{(n-1)N_{n-1}}, x_{nN_n})}_{\mathbf{x}_{N_1 N_2 \dots N_{(n-1)N_n}} \right\}. \quad (2) \end{aligned}$$

Utility functions: The set of utility functions, denoted $\mathcal{U} = \{u_i: \mathcal{A} \rightarrow \mathbb{R}, i = 1, \dots, n\}$, where u_i defines X_i ’s preferences over outcomes.

Since graph theory is an indispensable tool for the representation of social influence networks, we begin our development of CGT by reviewing the basic terminology and notation of graph theory that we will need.

Definition 2 A social influence network graph $G(\mathbf{X}, E)$ comprises a set of vertices $\mathbf{X} = \{X_1, \dots, X_n\}$ (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 among agents. The expression $X_i \rightarrow X_k$ means that influence propagates in only one direction — a directed edge from X_i to X_k . A path from X_i to X_k is a sequence of directed edges from X_i to X_k , denoted $X_i \mapsto X_k$. A path is a cycle, or closed path, for X_k if there is a path $X_k \mapsto X_k$.

A graph is said to be a directed acyclic graph, or DAG, if all edges are directed and there are no cycles. The parent set for X_i , denoted $\text{pa}(X_i) = \{X_{i_1}, \dots, X_{i_{q_i}}\}$, is a subset of \mathbf{X} such that $X_{i_k} \rightarrow X_i$, $k = 1, \dots, q_i$. If $\text{pa}(X_i) = \emptyset$ then X_i is a root vertex. The descendants of X_i , denoted $\text{de}(X_i) = \{X_k: X_i \mapsto X_k\}$, is the set of all vertices linked by directed paths from X_i .

We now introduce new terminology and definitions that apply to CGT.

Definition 3 The complementary set for X_i is the subset of all agents except X_i , denoted $X_{-i} = \mathbf{X} \setminus X_i$. The Cartesian product $\mathcal{A}_{-i} = \prod_{j \neq i} \mathcal{A}_j$ is termed the complementary action profile space.

A self-conjecture by X_i , denoted $X_i \models a_i$, is an element $a_i \in \mathcal{A}_i$ that is hypothesized by X_i as an action that she may choose.

A conditioning conjecture by X_i for X_j , denoted $X_j \models a_j$ for $j \neq i$, is an element $a_j \in \mathcal{A}_j$ that is hypothesized by X_i as an action that X_j may choose.

A conditioning conjecture profile, denoted $a_{-i} = (a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$, denoted $X_{-i} \models a_{-i}$, is a set of conditioning conjectures by X_i for X_{-i} .

A conjecture hypothesis is a hypothetical proposition with a conditioning conjecture profile as the antecedent and a self-conjecture as the consequent, denoted

$$\mathcal{H}_i(a_i|a_{-i}): X_{-i} \models a_{-i} \implies X_i \models a_i, \quad (3)$$

that is, if a_{-i} is a conditioning conjecture profile for X_{-i} , then a_i is a self-conjecture for X_i .

For each $a_{-i} \in \mathcal{A}_{-i}$, let $\succsim_{i|-i}$ denote a total preference ordering defined over $\mathcal{A}_i \times \mathcal{A}_i$. A conditional utility function is a mapping $\tilde{u}_{i|-i}: \mathcal{A}_i | \mathcal{A}_{-i} \rightarrow \mathbb{R}$, where the expression to the left of the conditioning symbol “|” is a self-conjecture for X_i and the expression to the right is a conditioning conjecture profile for X_{-i} , such that

$$\tilde{u}_{i|-i}(a_i|a_{-i}) \geq \tilde{u}_{i|-i}(a'_i|a_{-i}) \quad (4)$$

if

$$\mathcal{H}_i(a_i|a_{-i}) \succsim_{i|-i} \mathcal{H}_i(a'_i|a_{-i}), \quad (5)$$

meaning that the consequent $X_i \models a_i$ is either strictly preferred to the consequent $X_i \models a'_i$ or is indifferent, given the antecedent $X_{-i} \models a_{-i}$.¹

It is important to appreciate that, since a conjecture is merely a hypothesis and not a record of a fact, belief, or preference, the motivation for any conjecture hypothesis is application dependent. In particular, this conditional preference model enables agents to define their preference as a function of the preferences of others, in which case the expression $\tilde{u}_{i|jk}(a_i|a_j, a_k)$ corresponds to the narrative: “if X_j were to most prefer that a_j be realized and X_k were to most prefer that a_k be realized, then $\tilde{u}_{i|jk}(a_i|a_j, a_k)$ would be the utility to X_i if a_i were realized.”

Definition 4 A conditional game, denoted by the triple $\{\mathbf{X}, \mathcal{A}, \tilde{\mathcal{U}}\}$, is comprised of the following elements:

- A set of agents, $\mathbf{X} = \{X_1, \dots, X_n\}$, with each X_i possessing a finite action space $\mathcal{A}_i = \{x_{i1}, \dots, x_{iN_i}\}$, $i = 1, \dots, n$
- a set of conditional utility mass functions $\tilde{\mathcal{U}} = \{\tilde{u}_{i|-i}: \mathcal{A}_i | \mathcal{A}_{-i} \rightarrow [0, 1], i = 1, \dots, n\}$, such that²

$$\begin{aligned} \tilde{u}_{i|-i}(a_i|a_{-i}) &\geq 0 \quad \forall (a_i, a_{-i}) \in \mathcal{A}_i | \mathcal{A}_{-i} \\ \sum_{a_i} \tilde{u}_{i|-i}(a_i|a_{-i}) &= 1 \quad \forall a_{-i} \in \mathcal{A}_{-i}. \end{aligned} \quad (6)$$

¹We adopt the “tilde” notation to emphasize the distinction between standard (categorical) utilities, which are defined over outcomes, and conditional utilities, which are defined over actions given the actions of others.

²Without loss of generality we require conditional utility functions to be mass functions, which may require such normalization via positive affine transformations.

2.2 Hierarchical Networks

An important special case of a social influence network is a *hierarchical network*, which can be represented by a directed acyclic graph. Since there are no cycles, in such graphs the notions of parent and child are well defined, and the conditional utilities $\tilde{u}_{i|-i}$ may be expressed using the notation $\tilde{u}_{i|\text{pa}(i)}$. Let $\mathcal{A}_{\text{pa}(i)} = \mathcal{A}_{i_1} \times \cdots \times \mathcal{A}_{i_{q_i}}$ denote the Cartesian product of the domain of $\text{pa}(X_i)$ and let $\tilde{\alpha}_{\text{pa}(i)} = (a_{i_1}, \dots, a_{i_{q_i}}) \in \mathcal{A}_{\text{pa}(i)}$ denote the vector of conditioning conjectures for $\text{pa}(X_i)$.

Definition 5 A hierarchical conditional game is a conditional game with conditional utility functions of the form $\tilde{\mathcal{U}} = \{\tilde{u}_{i|\text{pa}(i)}: \mathcal{A}_i | \mathcal{A}_{\text{pa}(i)} \rightarrow [0, 1], i = 1, \dots, n\}$, where $\tilde{u}_{i|\text{pa}(i)}(a_i | \tilde{\alpha}_{\text{pa}(i)})$ is the utility function for $X_i \models a_i$ given that $\text{pa}(X_i) \models \tilde{\alpha}_{\text{pa}(i)}$.

The fact that conditional utilities are mass functions invites the application of probability theory to the theory of conditional games. The obvious isomorphic relationship between belief in an epistemological setting and preference in a praxeological setting enables the application of the syntax of probability theory to be applied to praxeological issues. Thus, a hierarchical network may be expressed using the mathematical syntax of probability theory with praxeological, rather than epistemological, semantics. The vertices of the graphical representation of such a network are analogous to random variables and the edges are analogous to conditional probability mass functions. Thus constituted, a hierarchical social influence network is analogous to a Bayesian network, where the vertices are random variables and the edges are conditional probability mass functions. To illustrate, consider a three-agent hierarchical network with graphical representation



where X_1 , a root vertex, possesses an unconditional utility $\tilde{u}_1: \mathcal{A}_1 \rightarrow [0, 1]$.

Given the isomorphic relationship between conditional game theory and probability theory, we may exploit a powerful result from Bayesian network theory.

Theorem 1 Let $\{X_1, \dots, X_n\}$ be a hierarchical social influence network with utility set $\tilde{\mathcal{U}}$. Then the coordination function is

$$\tilde{w}_{1:n}(a_1, \dots, a_n) = \prod_i \tilde{u}_{i|\text{pa}(i)}(a_i | \tilde{\alpha}_{\text{pa}(i)}). \tag{8}$$

If $\text{pa}(X_i) = \emptyset$ (i.e., X_i is a root vertex), then $\tilde{u}_{i|\text{pa}(i)} = \tilde{u}_i$, an unconditional mass function.

For a proof of this theorem and additional discussions of Bayesian networks, see Pearl (1988); Cowell et al. (1999); Lauritzen (1996); Jensen (2001); Sprites et al. (2000); Pearl (2009). In a conventional epistemological context, the coordination function would be the joint probability mass function, which characterizes all of the statistical dependency relationships that exist among a set of random variables as expressed via the conditional probability mass functions. In a praxeological context, the coordination function characterizes all of the social interdependency relationships that exist in a social influence network as expressed via the conditional utility mass functions.

This theorem establishes that, if influence relationships between neighboring vertices can be represented by conditional utility functions where the conditional dependencies flow in only one direction, then a unique joint mass function can be synthesized according to (8). In other words, when specifying dependency relationships with a directed acyclic graph, one need only be concerned with how the children are influenced by their parents, and not vice versa. Thus, all of the syntactical operations of probability theory apply to these functions, including marginalization, independence, and Bayes's rule.

3 Common Priors Through Markov Convergence

Adopting the hierarchical network structure allows us to appropriate the syntax of probability theory to model the social relationships that exist among a collective of socially connected agents that is analogous to the way in which the probability theory syntax models the statistical relationships that exist among a collective of statistically connected random variables. However, complying with this syntax restricts us to hierarchical models, since probability theory does not accommodate cyclic dependencies. Given the events A and B , the conditional probabilities $P(A|B)$ and $P(B|A)$ are related by Bayes's rule and, therefore, cannot be specified independently of each other. To put the matter another way, simultaneous feedback is not permitted. The most important and interesting strategic social relationships, however, involve feedback and reciprocal influence. Mindshaping can be modeled as a time-sequenced process of iterations with bidirectional flow. For such iterations to be useful, the propagation of conditional preferences through the network must ultimately result in actionable choices for each agent. Thus, the issue becomes one of convergence as cycles are recurrently traversed as time evolves. Convergence means that, in the limit, each individual will possess an unconditional preference ordering that may be used to identify solutions.

To address the convergence issue we again invoke the isomorphic relationship between conditional preferences and conditional beliefs as expressed via the syntax of probability theory, and exploit powerful results from stochastic process theory.

Definition 6 *A finite-state discrete-time stochastic process is a time-indexed set of random variables, denoted $\{Y(t), t = 0, 1, \dots\}$, which at any time can be in one of a finite number of states in the set $\mathcal{Y} = \{y_1, \dots, y_m\}$. The process is governed by a set of conditional probability measures that define the probability of transitioning from one state to another as time progresses. Let the sequence of events $\{Y(0) = s_0, Y(1) = s_1, \dots, Y(t) = s_t\}$ denote the path history of the process up to time t and let P denote a probability measure. The process satisfies the Markov condition if*

$$P[Y(t+1) = s_{t+1} | Y(0) = s_0, Y(1) = s_1, \dots, Y(t) = s_t] = P[Y(t+1) = s_{t+1} | Y(t) = s_t] \quad (9)$$

for all $(s_0, s_1, \dots, s_{t+1}) \in \times_{k=0}^{t+1} \mathcal{Y}_k$; that is, the probability of moving to any future state s_{t+1} at time $t+1$, given that it is in the state s_t at time t , is independent of the path taken to arrive at state s_t .³ In this sense, a Markov chain is memoryless, meaning that $Y(t-1)$ and $Y(t+1)$ are statistically conditionally independent, given $Y(t)$. Such a process is termed a finite state Markov chain.

³In the vernacular, the past and the future are independent given the present.

Since, by the structure of conditional mass functions, each agent's preferences are influenced by, and only by, her immediate ancestors, the network for a conditional game satisfies all of the conditions of a Markov chain. Thus, all of the methodology applicable to Markov chains is applicable to conditional games.

A two-agent scenario is a convenient way to introduce the study of cyclic games, but we must study games with more than two players in order to fully appreciate and deal with the complexity that quickly arises with the general case, where each agent influences all others and is influenced by all others. For reasons that will soon become apparent, we focus first on networks with a ring topology. Consider a directed ring graph $\{X_1, \dots, X_n\}$ of the form

$$\begin{array}{ccc}
 X_1 & \xrightarrow{\tilde{u}_{2|1}} & X_2 \\
 \tilde{u}_{1|n} \uparrow & & \downarrow \tilde{u}_{3|2} \\
 X_n & \leftarrow & X_3
 \end{array} \tag{10}$$

Let time t proceed in discrete increments, and let $X_i(t)$ denote X_i 's state at time t . Suppose it takes one unit of time to traverse one cycle $X_j(t) \rightarrow \dots \rightarrow X_j(t+1)$. Let $\delta = 1/(n+1)$ denote the time required to traverse from $X_j(j\delta)$ to $X_{j+1}((j+1)\delta)$ and consider the time-sequenced path

$$X_1(0) \xrightarrow{\tilde{u}_{2|1}} X_2(\delta) \xrightarrow{\tilde{u}_{3|2}} X_3(2\delta) \dashrightarrow X_n(n\delta) \xrightarrow{\tilde{u}_{1|n}} X_1(1) \dots, \tag{11}$$

which generates the closed path $X_1(0) \mapsto X_1(1)$. Starting at time $t = 0$, the coordination function corresponding to the segment $\{X_1(0) \rightarrow X_2(\delta)\}$ is, via Bayesian network theory (i.e., (8)):

$$\tilde{w}_{12}(a_1, a_2, 0) = \tilde{w}_1(a_1, 0) \tilde{u}_{2|1}(a_2|a_1), \tag{12}$$

where $\tilde{w}_1(a_{11}, 0)$ is the initial utility for X_1 at $t = 0$. The marginal utility for X_2 at time $t = \delta$ is

$$\hat{w}_2(a_2, \delta) = \sum_{a_1} \tilde{w}_{12}(a_1, a_2, \delta). \tag{13}$$

Now considering the segment $\{X_2(\delta) \rightarrow X_3(2\delta)\}$ (i.e., viewing $X_2(\delta)$ as a root vertex), and applying (8) and computing the marginal utility for X_3 at time $t = 2\delta$ yields

$$\begin{aligned}
 \tilde{w}_3(a_3, 2\delta) &= \sum_{a_2} \tilde{w}_{23}(a_2, a_3, 2\delta) \\
 &= \sum_{a_{22}} \tilde{w}_2(a_2, \delta) \tilde{u}_{3|2}(a_3|a_2).
 \end{aligned} \tag{14}$$

In general,

$$\tilde{w}_{i+1}(a_{i+1}, i\delta) = \sum_{a_i} \tilde{w}_i(a_i, (i-1)\delta) \tilde{u}_{i+1|i}(a_{i+1}|a_{ii}) \tag{15}$$

for $i = 1, 2, \dots, (\text{mod } n)$.

3.1 Matrix Form Dynamics Model

Matrix notation provides a convenient framework for modeling the dynamics of a sequence of acyclic segments of the form (11). For this development we will need to dispense with the practice of referring to elements of \mathcal{A}_i with the “arbitrary” notation a_i , and instead employ the specifically indexed notation for the members of \mathcal{A}_i as established in Definition 3.

To express the time-evolution of the network defined by (10) in matrix form, let

$$\tilde{\mathbf{w}}_i(i\delta) = \begin{bmatrix} \tilde{w}_i(x_{i1}, i\delta) \\ \vdots \\ \tilde{w}_i(x_{iN_i}, i\delta) \end{bmatrix} \quad (16)$$

denote the *utility mass vector* at time $t = i\delta$, let

$$T_{i+1|i} = \begin{bmatrix} \tilde{u}_{i+1|i}(x_{(i+1)1}|x_{i1}) & \cdots & \tilde{u}_{i+1|i}(x_{(i+1)1}|x_{iN_i}) \\ \vdots & & \vdots \\ \tilde{u}_{i+1|i}(x_{(i+1)N_{i+1}}|x_{i1}) & \cdots & \tilde{u}_{i+1|i}(x_{(i+1)N_{i+1}}|x_{iN_i}) \end{bmatrix} \quad (17)$$

denote the *agent-to-agent transition matrix*, and rewrite (15) as

$$\tilde{\mathbf{w}}_{i+1}(i\delta) = T_{i+1|i}\tilde{\mathbf{w}}_i((i-1)\delta) \quad (18)$$

for $i = 1, 2, \dots, \text{mod } k$. Expressing this cycle with the linkages represented by the transition matrices yields

$$\begin{array}{ccc} X_1 & \xrightarrow{T_{2|1}} & X_2 \\ T_{1|n} \uparrow & & \downarrow T_{3|2} \\ X_n & \leftarrow \cdots \leftarrow & X_3 \end{array} \quad (19)$$

Now define the *closed-loop transition matrices*

$$T_i = T_{i|i+n-1}T_{i+n-1|i+n-2} \cdots T_{i+2|i+1}T_{i+1|i} \pmod{n}. \quad (20)$$

After t cycles,

$$\tilde{\mathbf{w}}_i(t) = T_i\tilde{\mathbf{w}}_i(t-1) = T_iT_i\tilde{\mathbf{w}}_i(t-2) = \cdots = T_i^t\tilde{\mathbf{w}}_i(0). \quad (21)$$

Our goal is to investigate the behavior of these utility vectors as $t \rightarrow \infty$. To this end, we turn to the Markov chain convergence theorem.

Theorem 2 (Markov Chain Convergence) *Let T be a square matrix with nonnegative entries such that each column sums to unity and T is regular, meaning that there exists an integer m such that all elements of T^m are strictly positive. Then there exists a unique utility mass vector $\bar{\mathbf{w}}$ such that*

- $T\bar{\mathbf{w}} = \bar{\mathbf{w}}$, that is, $\bar{\mathbf{w}}$ is the eigenvector corresponding to the unique unit eigenvalue of T ;
- $\bar{T} = \lim_{t \rightarrow \infty} T^t = [\bar{\mathbf{w}} \ \cdots \ \bar{\mathbf{w}}]$;

- $\bar{\mathbf{w}} = \bar{T}\tilde{\mathbf{w}}(0)$ for every initial utility mass vector $\tilde{\mathbf{w}}(0)$.

This theorem establishes that $\bar{\mathbf{w}}$, the eigenvector corresponding to the unique unit eigenvalue of T , is the convergent utility mass vector. Thus, the conditional utilities of agents in a social influence network whose dynamic behavior is governed by transition matrices that meet the conditions of Theorem 2 will converge to a unique unconditional utility mass function, that is,

$$\bar{\mathbf{w}} = \lim_{t \rightarrow \infty} \tilde{\mathbf{w}}(t) = \begin{bmatrix} \bar{w}(\mathbf{x}_{i1}) \\ \vdots \\ \bar{w}(\mathbf{x}_{iN_i}) \end{bmatrix}. \quad (22)$$

Doob (1953) has established that convergence is exponentially fast.

3.2 Markov Equivalence

The above analysis applies to directed ring networks, but arbitrary networks do not conform to that special structure. In the limiting case of generality for a network, every agent is influenced by every other agent. To develop an approach to deal with this general structure it will be convenient to focus on a three-agent network, which provides sufficient complexity to identify the critical issues. Consider a network graph with bidirectional edges



where the edges are conditional utilities of the form $\tilde{u}_{i|jk}$ for $i|jk \in \{1|23, 2|31, 3|12\}$.⁴

We begin our study of this network by considering the acyclic graph fragment



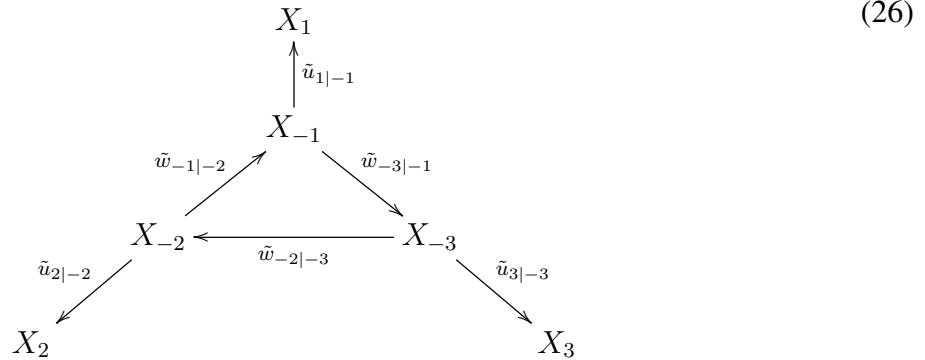
which qualifies as a Bayesian network with two root vertices. However, we cannot assume that these root vertices are independent, because this fragment is not isolated from the remainder of the network. Let us view $\{X_j, X_k\}$ as a dyadic root vertex yielding a graph fragment of the form

$$X_j X_k \xrightarrow{\tilde{u}_{i|jk}} X_i. \quad (25)$$

We now make an important observation: a graph of a network is *not* the network. It is a *representation* of the network, and representations are not unique. Of crucial interest is a representation comprising unidirectional edges that form a ring in order to apply Markov chain convergence theory.

⁴This indexing convention corresponds to a clockwise rotation around the network.

This structure motivates us to view the relationships between the complementary sets X_{-i} as the key to understanding the dynamics of this network. Accordingly, we introduce a *complementary network graph* to (23) of the form (assuming a clockwise rotation around the closed loop)



where the vertices are the complementary sets $X_{-i} = X_j X_k$ and the edges around the cycle are of the form $\tilde{w}_{-i|-(i+1)} \pmod{3} = \tilde{w}_{jk|ki}$, yet to be defined.⁵ We now have two graphical candidate representations of the network, and our issue is to relate them to each other. The key concept in this regard is *Markov equivalence*.

Definition 7 *Two graphs are Markov equivalent if they entail exactly the same conditional independence model (cf. Cowell et al. (1999); Jordan (2001)).*

Our goal is to define the conditional functions $\tilde{w}_{-(i-1)|-i}$ that map the social influence that the complementary subgroup X_{-i} exerts on the complementary subgroup $X_{-(i-1)}$ such that (23) and (26) are Markov equivalent. To proceed, we apply the chain rule of probability (the conditional product rule) to obtain the factorization

$$\tilde{w}_{ij|jk}(a_i, a_j|a'_j, a_k) = \tilde{w}_{j|ijk}(a_j|a_i, a'_j, a_k)\tilde{w}_{i|jk}(a_i|a'_j, a_k), \quad (27)$$

where a_j , the *conditioned* self-conjecture for X_j and a'_j , the *conditioning* self-conjecture for X_j , both independently range over \mathcal{A}_j . Notice, however, that $\tilde{w}_{j|ijk}(a_j|a_i, a'_j, a_k)$ contains a *self-conditioning* component; thus, it must be a degenerate mass function,⁶ yielding

$$\tilde{w}_{j|ijk}(a_j|a_i, a'_j, a_k) = \begin{cases} 1 & \text{if } a_j = a'_j \\ 0 & \text{otherwise.} \end{cases} \quad (28)$$

Substituting (28) into (27) yields

$$\tilde{w}_{ij|jk}(a_i, a_j|a'_j, a_k) = \begin{cases} \tilde{w}_{i|jk}(a_i|a_j, a_k) & \text{if } a_j = a'_j \\ 0 & \text{otherwise.} \end{cases} \quad (29)$$

⁵Although the vertices for the representation of a Bayesian network typically comprise single random variables with individual transition probabilities, there is nothing in the theory that prohibits a representation with the vertices comprising sets of random variables with joint transition probabilities. Similarly, a social influence network graph may be expressed with vertices comprising sets of agents with joint transition functions.

⁶This result follows from the definition of conditional probability: $P(A|A \wedge B) = P(A \wedge A \wedge B)/P(A \wedge B) = 1$.

Also, it is obvious that $\tilde{w}_{i|jk}(a_i|a'_j, a_k) = \tilde{u}_{i|jk}(a_i|a'_j, a_k)$, which yields

$$\tilde{w}_{i|jk}(a_i, a_j|a'_j, a_k) = \begin{cases} \tilde{u}_{i|jk}(a_i|a_j, a_k) & \text{if } a_j = a'_j \\ 0 & \text{otherwise} \end{cases} \quad (30)$$

for $ij|jk \in \{12|23, 23|31, 31|12\}$. With these specifications, the graphical representations (23) and (26) are Markov equivalent.

We illustrate this procedure with a 3×2 scenario. The transition matrix from $\{X_j, X_k\}$ to $\{X_i, X_j\}$ is

$$T_{ij|jk} = \begin{bmatrix} \tilde{w}_{i|jk}(x_{i1}, x_{j1}|x_{j1}, x_{k1}) & \tilde{w}_{i|jk}(x_{i1}, x_{j1}|x_{j1}, x_{k2}) \\ \tilde{w}_{i|jk}(x_{i1}, x_{j2}|x_{j1}, x_{k1}) & \tilde{w}_{i|jk}(x_{i1}, x_{j2}|x_{j1}, x_{k2}) \\ \tilde{w}_{i|jk}(x_{i2}, x_{j1}|x_{j1}, x_{k1}) & \tilde{w}_{i|jk}(x_{i2}, x_{j1}|x_{j1}, x_{k2}) \\ \tilde{w}_{i|jk}(x_{i2}, x_{j2}|x_{j1}, x_{k1}) & \tilde{w}_{i|jk}(x_{i2}, x_{j2}|x_{j1}, x_{k2}) \\ \tilde{w}_{i|jk}(x_{i1}, x_{j1}|x_{j2}, x_{k1}) & \tilde{w}_{i|jk}(x_{i1}, x_{j1}|x_{j2}, x_{k2}) \\ \tilde{w}_{i|jk}(x_{i1}, x_{j2}|x_{j2}, x_{k1}) & \tilde{w}_{i|jk}(x_{i1}, x_{j2}|x_{j2}, x_{k2}) \\ \tilde{w}_{i|jk}(x_{i2}, x_{j1}|x_{j2}, x_{k1}) & \tilde{w}_{i|jk}(x_{i2}, x_{j1}|x_{j2}, x_{k2}) \\ \tilde{w}_{i|jk}(x_{i2}, x_{j2}|x_{j2}, x_{k1}) & \tilde{w}_{i|jk}(x_{i2}, x_{j2}|x_{j2}, x_{k2}) \end{bmatrix} \quad (31)$$

Substituting (30) into (31) yields

$$T_{ij|jk} = \begin{bmatrix} \tilde{u}_{i|jk}(x_{i1}|x_{j1}, x_{k1}) & \tilde{u}_{i|jk}(x_{i1}|x_{j1}, x_{k2}) & 0 & 0 \\ 0 & 0 & \tilde{u}_{i|jk}(x_{i1}|x_{j2}, x_{k1}) & \tilde{u}_{i|jk}(x_{i1}|x_{j2}, x_{k2}) \\ \tilde{u}_{i|jk}(x_{i2}|x_{j1}, x_{k1}) & \tilde{u}_{i|jk}(x_{i2}|x_{j1}, x_{k2}) & 0 & 0 \\ 0 & 0 & \tilde{u}_{i|jk}(x_{i2}|x_{j2}, x_{k1}) & \tilde{u}_{i|jk}(x_{i2}|x_{j2}, x_{k2}) \end{bmatrix} \quad (32)$$

The closed-loop transition matrices are

$$T_{ij} = T_{ij|jk}T_{jk|ki}T_{ki|ij} \quad (33)$$

for $ij \in \{12, 23, 31\}$.

We are now in a position to apply the Markov chain convergence theorem to the ring with vertices comprising the dyads $\{X_i X_j\}$, with the eigenvectors corresponding to the unique unity eigenvalues of these matrices comprising the *steady-state complementary correlation functions* for $\{X_i, X_j\}$, denoted

$$\bar{\mathbf{w}}_{-i} = \bar{\mathbf{w}}_{jk} = \begin{bmatrix} \bar{w}_{jk}(x_{j1}, x_{j1}) \\ \bar{w}_{jk}(x_{j1}, x_{j2}) \\ \bar{w}_{jk}(x_{j2}, x_{j1}) \\ \bar{w}_{jk}(x_{j2}, x_{j2}) \end{bmatrix}. \quad (34)$$

As discussed by Aumann (1987), the notion of *correlation* in game theory arises via a signal generated by some randomizing device that generates information for each player that enables them to identify *correlated* strategies. For our scenario, the diffusion of conditional utilities throughout the network generates a signal to the players that enables them to define the functions \bar{w}_{-i} that enable them to achieve *praxeological correlation* analogous to the *epistemological correlation* in Aumann's setting.

The individual correlated utility vectors are computed as

$$\bar{\mathbf{w}}_i = \begin{bmatrix} \bar{w}_i(x_{i1}) \\ \bar{w}_i(x_{i2}) \end{bmatrix} = T_{i|jk} \bar{\mathbf{w}}_{jk}. \quad (35)$$

with

$$T_{i|jk} = \begin{bmatrix} \tilde{u}_{i|jk}(x_{i1}|x_{j1}, x_{k1}) & \tilde{u}_{i|jk}(x_{i1}|x_{j1}, x_{k2}) \\ \tilde{u}_{i|jk}(x_{i2}|x_{j1}, x_{k1}) & \tilde{u}_{i|jk}(x_{i2}|x_{j1}, x_{k2}) \\ \tilde{u}_{i|jk}(x_{i1}|x_{j2}, x_{k1}) & \tilde{u}_{i|jk}(x_{i1}|x_{j2}, x_{k2}) \\ \tilde{u}_{i|jk}(x_{i2}|x_{j2}, x_{k1}) & \tilde{u}_{i|jk}(x_{i2}|x_{j2}, x_{k2}) \end{bmatrix}. \quad (36)$$

3.3 General n th Order Networks

The methodology developed for a general three-agent network may be extended to an arbitrary n th order network as follows.

1. Generate the ring graph for the complementary network representation

$$\begin{array}{ccc} X_{-n} & \xrightarrow{\tilde{w}_{-1|-n}} & X_{-1} \\ \tilde{w}_{-n|-(n-1)} \uparrow & & \downarrow \tilde{w}_{-2|-1} \\ X_{-(n-1)} & \leftarrow & X_{-2} \end{array} \quad (37)$$

with vertices comprising the complementary subgroups X_{-i} and edges comprising the conditional complementary correlation functions $\tilde{w}_{-i|-(i+1)}(a_{-i}|a_{-(i+1)}) \pmod{n}$:

$$\begin{aligned} \tilde{w}_{-1|-n}(a_{-1}|a_{-n}) &= \tilde{w}_{2:n|1:(n-1)}(a_2, \dots, a_n | a'_1, \dots, a'_{n-1}) \\ \tilde{w}_{-2|-1}(a_{-2}|a_{-1}) &= \tilde{w}_{3:1|2:n}(a_3, \dots, a_n, a_1 | a'_2, \dots, a'_n) \\ &\vdots \\ \tilde{w}_{-n|-(n-1)}(a_{-n}|a_{-(n-1)}) &= \tilde{w}_{1:(n-1)|n:(n-2)}(a_1, \dots, a_{n-1} | a'_n, a'_1, \dots, a'_{n-2}) \end{aligned} \quad (38)$$

where the notation $2 : n := 2, 3, \dots, n$ and so forth, and indexing is \pmod{n} ; that is, $3 : (n+1) \pmod{n} := 3 : 1$ and so forth.

2. Invoke Markov equivalence to define the conditional complementary correlation functions

$$\tilde{w}_{-1|-n}(a_2, \dots, a_n | a'_1, \dots, a'_{n-1}) = \begin{cases} \tilde{u}_{1|-1}(a_1 | a_2, \dots, a_n) & \text{if } (a_2, \dots, a_{n-1}) = (a'_2, \dots, a'_{n-1}) \\ 0 & \text{otherwise,} \end{cases} \quad (39)$$

$$\tilde{w}_{-2|-1}(a_3, \dots, a_n, a_1 | a'_2, \dots, a'_n) = \begin{cases} \tilde{u}_{2|-2}(a_2 | a_3, \dots, a_n, a_1) & \text{if } (a_3, \dots, a_n) = (a'_3, \dots, a'_n) \\ 0 & \text{otherwise,} \end{cases} \quad (40)$$

continuing,

$$\tilde{w}_{-n|(n-1)}(a_1, \dots, a_{n-1} | a'_n, a'_1, \dots, a'_{n-1}) = \begin{cases} \tilde{u}_{n|-n}(a_n | a_1, \dots, a_{n-1}) & \text{if } (a_1, \dots, a_{n-2}) = (a'_1, \dots, a'_{n-2}) \\ 0 & \text{otherwise,} \end{cases} \quad (41)$$

which are used to populate the *conditional complementary transition matrices* $T_{-i|-(i+1)} \pmod{n}$.

3. Compute the closed-loop transition matrices

$$T_{-i} = T_{-i|-(i+1)} T_{-(i+1)|-(i+2)} \cdots T_{-(i-1)|-i} \pmod{n} \quad (42)$$

and apply the Markov chain convergence theorem to generate the complementary subgroup correlation functions \bar{w}_{-i} as the eigenvectors corresponding to the unique unit eigenvalues of T_{-i} .

4. Populate the conditional transition matrices $T_{i|-i}$; all arguments and subscripts are indexed \pmod{n} :

$$T_{i|-i} = \begin{bmatrix} \tilde{u}_{i|-i}(x_{i1} | x_{(i+1)1}, \dots, x_{(i+n)1}) & \cdots \\ \tilde{u}_{i|-i}(x_{i2} | x_{(i+1)1}, \dots, x_{(i+n)1}) & \cdots \\ \vdots & \\ \tilde{u}_{i|-i}(x_{iN_i} | x_{(i+1)1}, \dots, x_{(i+n)1}) & \cdots \\ \tilde{u}_{i|-i}(x_{i1} | x_{(i+1)N_{i+1}}, \dots, x_{(i+n)N_{i+n}}) \\ \tilde{u}_{i|-i}(x_{i2} | x_{(i+1)N_{i+1}}, \dots, x_{(i+n)N_{i+n}}) \\ \vdots \\ \tilde{u}_{i|-i}(x_{iN_i} | x_{(i+1)N_{i+1}}, \dots, x_{(i+n)N_{i+n}}) \end{bmatrix} \quad (43)$$

and compute the individual *correlated utilities*

$$\bar{w}_i = T_{i|-i} \bar{w}_{-i}. \quad (44)$$

3.4 The Harsanyi Doctrine

In standard game theory, various technologies are used to model processes by which agents can learn about other agents' utility functions by observing actions (Fudenberg and Levine, 1998). Such learning is powerful if Bayesian agents (who know that other players are Bayesians) play only pure strategies. But they fail under a wide range of common assumptions if players are assumed to play mixed strategies, as are recommended in many games by Nash equilibrium. This is the problem space in which Aumann (1987) demonstrates the major result that if players have common knowledge that all players are Bayesian expected utility maximizers (i.e., conform to the axioms of Savage (1954), and if they have common priors, then they can identify correlated equilibria in pure strategies.

As we will discuss in the next section, the power of this result for empirical application to human strategic interactions is threatened by the finding that conformity to expected utility theory

(EUT) is observed only in a minority of people. In this section, we first focus on the other condition supporting correlated equilibrium as a general solution to equilibrium selection, the establishment of common priors. The idea that all Bayesian players in games should have common priors is often referred to as “the Harsanyi Doctrine”, after Harsanyi’s (1977) appeal to it in his general theory of strategic bargaining.

Aumann’s (1987) defense of the Harsanyi Doctrine may reasonably be called ‘philosophical’, in that it consists mainly in establishing clearly what it *means*. As Aumann stresses, it does not mean that all players in a game assign the same subjective probabilities to events. It means rather that any differences among them should be fully explained by differences in the information available to them. In particular, players agree on the probability measures on possible states of the world relevant to their assessments of utilities associated with outcomes of games in which they jointly participate.

Aumann does not argue that such agreement is observed as an empirical matter. He argues that it is a presupposition of applications of game theory that depend on “Bayesian rationality”; indeed it is a presupposition of the ‘economic point of view’ more generally. To distance ourselves from the dominant tradition in philosophical decision theory (e.g. Bradley (2017)) that regards general ‘rationality’ as a specifiable target virtue for both decision makers and cognitive inquirers, we prefer to express Aumann’s point in different language. (For motivations see Ross (2023).) Economic models are populated by ‘agents’ in a technical sense. Agents seek to optimize their utility given constraints. In games these constraints include actions available to other agents. Simultaneous specification of these constraints by interacting agents, i.e., players, requires them to include one another’s estimates of one another’s probabilities of actions in their priors. Then if all players know that all players are EUT optimizers, then all of these priors should be the same. If they then receive different information, then of course their posteriors may differ. But if they receive common signals, and know that they do, then even if they also know that others have received varying private signals, they can identify correlated equilibria.

It makes no difference to this reasoning whether players make conjectures about actions conditional on outcomes, as in standard game theory, or on actions, as in CGT. Indeed, common priors are implicit in the Markov convergence process from Section 3.3 above by which players of conditional games identify the transition matrices (43). Using that information, each X_i is able to generate her complementary correlation function \bar{w}_{-i} and apply (44) to generate her correlated individual utility function \bar{w}_i . Given that information, she can compute

$$T_{-j|-i} = T_{-j|-(j+1)}T_{-(j+1)|-(j+2)} \cdots T_{-(j-1)|-i} \pmod{n} \quad (45)$$

to obtain

$$\bar{w}_{-j} = T_{-j|-i}\bar{w}_{-i} \quad (46)$$

from which she can generate X_j ’s correlated utility function via

$$\bar{w}_j = T_{j|-j}\bar{w}_{-j}. \quad (47)$$

Thus, each player has knowledge of each other’s unconditional utility function as well as her own. These utilities may then be juxtaposed into a payoff array for which correlated equilibria may be identified. Figure 1 displays the correlated payoff matrix for a 2×2 conditional game

CGT models a mindshaping process. It should hardly come as a shock to intuitions that mindshaping supplies agents in a social influence network with shared signals that support correlation.

Table 1: The correlated payoff matrix for a 2×2 conditional game.

		X_2	
		x_{21}	x_{22}
X_1	x_{11}	$(\bar{w}_1(x_{11}), \bar{w}_2(x_{21}))$	$(\bar{w}_1(x_{11}), \bar{w}_2(x_{22}))$
	x_{12}	$(\bar{w}_1(x_{12}), \bar{w}_2(x_{21}))$	$(\bar{w}_1(x_{12}), \bar{w}_2(x_{22}))$

That is one way, true to Zawidzki’s motivations and evidence, of stating the very point of mind-shaping.

As noted, however, all of this reasoning has assumed that agents are expected utility maximizers. Complications for application evidently loom if and when this assumption is in doubt. We will now argue that CGT furnishes special resources for dealing with this complication.

4 Accommodating Rank Dependent Utility

Aumann’s argument for correlated equilibrium as a general solution to the equilibrium selection problem in game theory applies to agents who are expected utility maximizers, that is, whose choices respect the axioms of Savage (1954). Aumann’s paper is about theory, not methodology for the empirical study of choice. Therefore, it is not a criticism of his argument to observe that laboratory evidence indicates that most people, at least outside of institutional contexts that incentivize and equip them to compute expected utility and act on the basis of such computations, tend statistically to violate two Savage axioms: independence, and reduction of compound lotteries (ROCL) (Bourgeois-Gironde, 2020; Harrison et al., 2015). Most of the relevant evidence from economic experiments had in any event not yet been produced when Aumann wrote his 1987 paper.

What *is* interesting here, even to the student of purely general theory, is that the family of models that experimental economists use to generalize EUT (in the sense that the family formally nests it) and to accommodate relaxation of the behaviorally neglected axioms incorporates a feature that Aumann (1987) considers at some length: so-called ‘personal probabilities’. We will first indicate why this concept arises in Aumann’s discussion and what he says about it.

As Aumann reconstructs the Harsanyi Doctrine, it amounts to the denial that agents in games can coherently be thought of as having personal probabilities, even though in Savage’s actual treatment probabilities are derived individually from each agent’s choices. Aumann notes that economists as of the time of his writing had occasionally visited the implications of personal probabilities, but that such investigations had found no “considerable echo”. “Apparently,” Aumann suggests, “economists feel that this kind of analysis is too inconclusive for practical use, and side-steps the major economic issues” (Aumann, 1987, 15). A couple of pages earlier, he had offered philosophical reflections to support this attitude, according to which utilities in the Savage setting are rightly and crucially personal, but beliefs about probabilities should be restricted in accordance with the Harsanyi Doctrine:

... utilities directly express tastes, which are inherently personal. It would be silly to

talk about “impersonal tastes”, tastes that are “objective” or “unbiased.” But it is not at all silly to talk about unbiased probability estimates, and even to strive to achieve them. On the contrary, people are often criticized for wishful thinking - for letting their preferences color their judgment. One cannot sensibly ask for expert advice on what one’s tastes should be; but one may well ask for expert advice on probabilities (Aumann, 1987, 13).

This intuitive asymmetry currently makes itself felt in arguments about how to do “behavioral welfare economics” (see Harrison and Ross (2018, 2023) for general discussions). The general family of utility models that best capture the statistical majority of observed choices under risk in the economic laboratory (Harrison and Ross, 2016) are rank-dependent utility (RDU) models following Quiggin (1982, 1993). We technically outline the relevant features of RDU below. For the moment, what is crucial is that RDU attaches decision weights to probabilities that reflect utility rankings. Thus Savage’s postulate that subjective probabilities do not depend on utilities is violated in RDU. It would be rash to simply assert that these are personal probabilities in the sense Aumann intends. But that that is a natural interpretation is indicated by divided opinion in welfare economics over whether decision weights should or should not be over-ruled by a policy advisor for agents whose choices reveal RDU structure. In terms of the quotation from Aumann above, should such revealed preferences be adjusted away by the advisor because their use can be “criticized”? Or does this amount to paternalistically over-ruling the client agent’s preferences? As Harrison and Ross (2018) point out, there can be sound justifications for choosing on the basis of the “fat tails” in distributions associated with typically observed decision weights. Perhaps the agent assumes that many of her strategic interactions occur against a background of interested manipulation by designers of rules and institutions. More prosaically, perhaps she knows that when she estimates distributions, she typically has relatively sparse observations from tails, so she might prudently operate a heuristic to correct for risky consequences of this sparse information. Buchak (2013), in pursuing the philosophical project from which we distanced ourselves in Section 3, argues that RDU is the best decision model for risky choices that a “generally rational” agent should adopt, in preference to EUT. So she interprets RDU decision weights as personal probabilities in the strongest conceivable sense.

In the remainder of the paper, we argue that CGT naturally provides a mechanism for pushing this philosophical dispute into the background, in a way that allows economists to get on with business without needing to take a stand on it. More importantly, our argument can dissolve the practical tension between Aumann’s general theoretical conclusion and empirical application of game theory to subjects whose actual choices are best described by RDU models. The basic intuition behind the argument, on which we will elaborate in the concluding discussion, is that the interpretative ambiguity around the normative status of decision weights arises from the fact that they reflect social-strategic influence and experience. Mindshaping influences both beliefs and preferences *jointly*. Thus at the level of theory, the CGT stage of analysis, which captures mindshaping elements, is the place to capture rank-dependence. At the next stage of analysis, where standard game theory is applied to unconditional utilities generated by the conditional Markov process, effects of decision weights are then already reflected in the transition matrices that constitute the shared signal on which correlated equilibria are identified. At that stage, in effect, agents should be modeled under the full set of Savage axioms.

Before we proceed, we should note that this has no practical implications for the experimental

economist's identification or estimation of her empirical model. But that economist might well want to compare game outcomes she observes in the lab or field with what abstract game theory tells her to expect. If her subjects found equilibrium, she might wonder how they did so, and speculate about psychological mechanisms. Our reflections suggest that she should consider mindshaping mechanisms, some of which might have operated before the experiment started.

4.1 Generalized Expected Utility

As noted previously, a specification of an RDU model generalizes EUT, Savage's theory of choice under uncertainty: RDU formally nests EUT. We therefore characterize RDU beginning from EUT. The elements are a) a finite set \mathcal{S} of states of the world, b) a finite set \mathcal{C} of consequences,⁷ c) a set \mathcal{F} of action functions (random variables) of the form $f: \mathcal{S} \rightarrow \mathcal{C}$ such that $f(s)$ determines the consequence in \mathcal{C} if f is adopted when $s \in \mathcal{S}$ is realized, and d) a total preference ordering \succsim defined over $\mathcal{F} \times \mathcal{F}$.

Savage's theory gives us a manual for evaluating the elements of \mathcal{F} when there is uncertainty regarding the state of the world. The key existence result is that, if the Savage (1954) postulates hold, then there is a discrete probability measure P defined over $2^{\mathcal{S}}$, the power set of \mathcal{S} , and a unique (up to a positive affine transform) utility assignment $u: \mathcal{C} \rightarrow \mathbb{R}$ such that, for $f, f' \in \mathcal{F}$,

$$f \succsim f' \text{ if, and only if, } E[u(f)] \geq E[u(f')], \quad (48)$$

where

$$E[u(f)] = \sum_{c \in \mathcal{C}} u(c) P[f^{-1}(c)] \quad (49)$$

is the *expected utility* of f . The set $f^{-1}(c) = \{s \in \mathcal{S}: f(s) = c\}$ is the inverse image of $c \in \mathcal{C}$ under the mapping $f: \mathcal{S} \rightarrow \mathcal{C}$ and

$$P[f^{-1}(c)] = \sum_{s \in f^{-1}(c)} p(s), \quad (50)$$

where $p(s) = P(\{s\})$ (where $\{s\}$ denotes a singleton set) is the probability mass function corresponding to the probability measure P .⁸

Represent an expected payoff to an agent considering a risky choice as $v: \mathcal{C} \rightarrow \mathbb{R}$. Her deliberations are governed by two attributes: a) the marginal utility associated with receiving $v(c)$ and b) her attitude associated with its receipt in the presence of risk. These attributes are encoded with a utility operator U defined over the set of payoff functions that generates a utility function $u: \mathcal{C} \rightarrow \mathbb{R}$ such that $u(c) = U[v(c)]$. A well-known risk response function, which we use for convenience, is the constant relative risk aversion (CRRA) utility function, which has the general form

$$CRRA(\tilde{v}) = \begin{cases} \frac{\tilde{v}^{1-r}-1}{1-r} & r \neq 1 \\ \log(\tilde{v}) & r = 1 \end{cases}. \quad (51)$$

⁷Savage's theory extends to scenarios with infinitely many consequences, but here we restrict treatment to the finite case.

⁸The inverse image of $f: \mathcal{A} \rightarrow \mathcal{A}$ is a mapping $f^{-1}: 2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$, that is, f^{-1} a mapping of the power set of \mathcal{A} to itself. Precise notation, then requires that we write $c \in \mathcal{A}$, $\{c\} \subset \mathcal{A}$, and $\{c\} \in 2^{\mathcal{A}}$. Thus, to be notationally precise, we would write $f^{-1}(\{c\})$ rather than $f^{-1}(c)$, where $\{c\}$ is a singleton set. However, since, for our application, we are only interested in the inverse image of singleton sets, we slightly abuse notation and eliminate the braces. Unless f is an injective mapping (also termed a one-to-one mapping), the inverse image $f^{-1}(c)$ may contain more than one element of \mathcal{A} .

The parameter r is termed the coefficient of constant relative risk aversion: $r = 0$ corresponds to risk neutrality, $r < 0$ to a risk seeking attitude, and $r > 0$ to risk aversion. Since conditional game theory requires utilities to take values in the unit interval, we perform a positive affine transform

$$\tilde{u} = U(\tilde{v}; r) = (1 - r) CRRA(\tilde{v}) + 1 = \tilde{v}^{1-r}, r \leq 1 \Rightarrow \begin{cases} \text{concave} & 0 < r < 1 \\ \text{linear} & r = 0 \\ \text{convex} & r < 0 \end{cases} \quad (52)$$

(we require $r < 1$ to ensure that the transform is positive)

We now define rank-dependent utility as developed by Quiggin (1982). Whereas Savage expected utility is defined as the sum of the products of the utilities of the outcomes and the probabilities of their realizations, rank-dependent expected utility is defined as the sum of the products of the utilities of the outcomes and decision weights ascribed to the outcomes that are functions of their ranking positions as well as their probabilities, thereby violating Savage's postulate that the probability cannot depend on the utility. Quiggin defines what we will call *Quiggin operators*, which transform probabilities into rank-dependent decision weights. We will refer to the application of the Quiggin operators as *Quigginization*. The procedure is as follows. Let $\{c_1, \dots, c_N\}$ denote the outcomes ordered from most preferred to least preferred, yielding

$$c_1 \succ c_2 \succ \dots \succ c_N, \quad (53)$$

with associated probabilities

$$P[f^{-1}(c_1)], \dots, P[f^{-1}(c_N)]. \quad (54)$$

Let $\omega: [0, 1] \rightarrow [0, 1]$ be a strictly increasing and continuous *probability weighting function*, with $\omega(0) = 0$ and $\omega(1) = 1$. The Quiggin operators produce the functions on the right hand side of the following expressions:

$$\begin{aligned} \pi(c_1) &= \omega[P[f^{-1}(c_1)]] \\ \pi(c_2) &= \omega[P[f^{-1}(c_1)] + P[f^{-1}(c_2)]] - \omega[P[f^{-1}(c_1)]] \\ &\vdots \\ \pi(c_k) &= \omega[P[f^{-1}(c_1)] + \dots + P[f^{-1}(c_k)]] - \omega[P[f^{-1}(c_1)] + \dots + P[f^{-1}(c_{k-1})]] \\ &\vdots \\ \pi(c_N) &= 1 - \omega[P[f^{-1}(c_1)] + \dots + P[f^{-1}(c_{N-1})]]. \end{aligned} \quad (55)$$

The rank-dependent utility is thus defined as

$$RDU[u] = \sum_k u(c_k) \pi(c_k). \quad (56)$$

A probability weighting function that has gained wide acceptance due to its flexibility for representing empirical choice data is due to Prelec (1998):

$$\omega(p) = P(p) \quad (57)$$

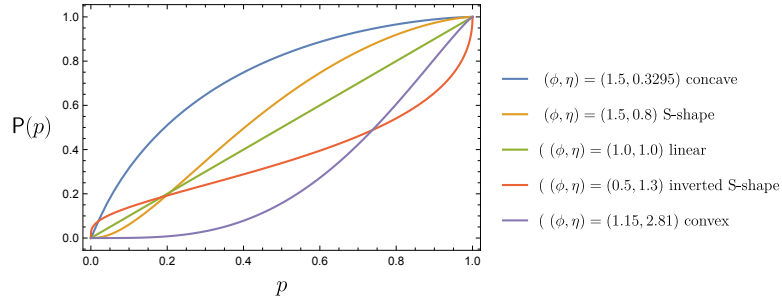


Figure 1: Prelec operator P .

where P is the Prelec operator defined by

$$P(p) = \exp \left[-\eta [-\log p]^\phi \right], \quad \phi > 0, \quad \eta > 0. \quad (58)$$

The parameters (ϕ, η) can be chosen to generate different shapes (e.g., convex, concave, S-shaped, and inverse S) that often provide best fits to data. Figure 1 displays the Prelec operator for various values of (ϕ, η) .

4.2 RDU in Conditional Games

We extend the model of generalized expected utility, which applies originally to individual decision making under risk, to multiple agent interactive decision scenarios. The state of the world for X_i is the element $s_i \in \mathcal{A}_i$ under consideration by X_i as the action to be chosen. A consequence for X_i , denoted c_i , is the element $c_i \in \mathcal{A}_i$ that is realized.

Let $\tilde{\mathcal{V}} = \{\tilde{v}_{i|-i}: \mathcal{A}_i | \mathcal{A}_{-i} \rightarrow [0, 1], i = 1, \dots, n\}$ denote a set of conditional payoffs. Let $\tilde{\mathcal{F}} = \{\tilde{f}_1, \dots, \tilde{f}_n\}$ denote a set of action functions $\tilde{f}_i: \mathcal{A}_i \rightarrow \mathcal{A}_i$ such that $c_i = \tilde{f}_i(s_i)$ is the consequence to X_i when she in takes action $s_i \in \mathcal{A}_i$. Let $\tilde{\mathcal{P}} = \{\tilde{p}_{i|-i}: \mathcal{A}_i | \mathcal{A}_{-i} \rightarrow [0, 1], i = 1, \dots, n\}$ denote a set of conditional probability mass functions defined over X_i 's state of the world such that $\tilde{p}_{i|-i}(s_i | \tilde{\alpha}_i)$ is the probability of taking action $s_i \in \mathcal{A}_i$ given that X_{-i} is in world state $\tilde{\alpha}_i \in \mathcal{A}_{-i}$.

Let U_i denote a utility operator for X_i . A straightforward procedure would be to apply U_i to each conditional payoff. This transformation, however, does not preserve the requirement that the utility must be a mass function as required by CGT, since, in general, $\sum_{c_j} U_i[\tilde{v}_{i|-i}(c_j | c_{-i})] \neq 1$. To address this issue, we turn to the procedure introduced by Quiggin to transform a probability mass function via the Quiggin operators, defined by (55), into a set of rank-dependent decision weights that have the syntactical structure of a probability mass function (they are nonnegative and sum to unity). In decision theory under RDU, these decision weights are understood as reflecting idiosyncratic *beliefs* about probabilities of outcomes based on their utility ranking, and the decision weights account for this epistemological equivocation. In the context of conditional game theory, we exploit a structural analogy between these decision weights and praxeological rank-dependent utility weights that reflect the idea that an agent might strategically adjust her preferences expressed by actual *or* conjectured choices to reflect uncertainty about the behavior of those with whom she

interacts. Since the utilities of conditional game theory comply with the probability syntax, we may define a *utility weighting function* by transforming conditional utilities to account for this praxeological uncertainty via the Quiggin operators defined by (61).

To proceed, we establish the preference ordering for each conditioning self-conjecture subprofile in \mathcal{A}_{-i} where, expressed in lexicographical order (mod n), the sets of complementary subprofiles are

$$\begin{aligned}
\mathcal{A}_{-1} &= \left\{ \underbrace{(x_{21}, x_{31}, \dots, x_{(n-1)1}, x_{n1})}_{\tilde{\alpha}_{11}}, \underbrace{(x_{21}, x_{31}, \dots, x_{(n-1)1}, x_{n2})}_{\tilde{\alpha}_{12}}, \right. \\
&\quad \left. \dots, \underbrace{(x_{2N_2}, x_{3N_3}, \dots, x_{(n-2)N_{(n-2)}}, x_{nN_n})}_{\tilde{\alpha}_{1N-1}} \right\} \\
\mathcal{A}_{-2} &= \left\{ \underbrace{(x_{31}, x_{41}, \dots, x_{n1}, x_{11})}_{\tilde{\alpha}_{21}}, \underbrace{(x_{31}, x_{41}, \dots, x_{n1}, x_{12})}_{\tilde{\alpha}_{22}}, \right. \\
&\quad \left. \dots, \underbrace{(x_{3N_3}, x_{4N_4}, \dots, x_{nN_n}, x_{1N_1})}_{\tilde{\alpha}_{2N-2}} \right\} \\
&\quad \vdots \\
\mathcal{A}_{-n} &= \left\{ \underbrace{(x_{11}, x_{21}, \dots, x_{(n-2)1}, x_{(n-1)1})}_{\tilde{\alpha}_{n1}}, \underbrace{(x_{11}, x_{21}, \dots, x_{(n-2)N_{(n-2)}}, x_{(n-1)2})}_{\tilde{\alpha}_{n2}}, \right. \\
&\quad \left. \dots, \underbrace{(x_{1N_1}, x_{2N_2}, \dots, x_{(n-2)N_{(n-2)}}, x_{(n-1)N_{(n-1)}})}_{\tilde{\alpha}_{nN-n}} \right\}
\end{aligned} \tag{59}$$

with the cardinality of \mathcal{A}_{-i} is $N_{-i} = \prod_{j \neq i} N_j$. For each $\tilde{\alpha}_{ik} \in \mathcal{A}_{-i}$, let $\{j_{ik_1}, \dots, j_{ik_{N_i}}\}$ be a permutation of $\{1, \dots, N_{-i}\}$ such that the conditional payoffs are ordered according to decreasing preference, that is,

$$\tilde{v}_{i|-i}(x_{ij_{k_1}} | \tilde{\alpha}_{ik}) \geq \dots \geq \tilde{v}_{i|-i}(x_{ij_{k_{N_{-i}}}} | \tilde{\alpha}_{ik}). \tag{60}$$

We apply Quigginization to define the *rank-dependent conditional utilities*

$$\begin{aligned}
\tilde{u}_{i|-i}(x_{ij_{k_1}} | \tilde{\alpha}_{ik}) &= \mathbf{U}_i[\tilde{v}_{i|-i}(x_{ij_{k_1}} | \tilde{\alpha}_{ik})] \\
\tilde{u}_{i|-i}(x_{ij_{k_2}} | \tilde{\alpha}_{ik}) &= \mathbf{U}_i[\tilde{v}_{i|-i}(x_{ij_{k_1}} | \tilde{\alpha}_{ik}) + \tilde{v}_{i|-i}(x_{ij_{k_2}} | \tilde{\alpha}_{ik})] - \mathbf{U}_i[\tilde{v}_{i|-i}(x_{ij_{k_1}} | \tilde{\alpha}_{ik})] \\
&\quad \vdots \\
\tilde{u}_{i|-i}(x_{ij_{k_{N_{-i}}}} | \tilde{\alpha}_{ik}) &= 1 - \mathbf{U}_i[\tilde{v}_{i|-i}(x_{ij_{k_1}} | \tilde{\alpha}_{ik}) + \dots + \tilde{v}_{i|-i}(x_{ij_{k_{(N_{-i}-1)}}} | \tilde{\alpha}_{ik})].
\end{aligned} \tag{61}$$

By construction, $\tilde{u}_{i|-i}$ is a mass function, that is,

$$\begin{aligned}
\tilde{u}_{i|-i}(c_i | \tilde{\alpha}_{ik}) &\geq 0 \quad \forall \tilde{\alpha}_{ik} \in \mathcal{A}_{-i} \\
\sum_j \tilde{u}_{i|-i}(c_j | \tilde{\alpha}_{ik}) &= 1.
\end{aligned} \tag{62}$$

As these conditional utilities propagate according to the Bayesian network/Markov convergence theory, they generate correlated payoffs \bar{w}_i as defined by (44). The outcome c_i for each

agent taking action is governed by the conditional probability mass functions $\tilde{p}_{i|-i}$, which must also be propagated through the network to generate a corresponding probability mass function \bar{p}_i , which governs the outcome when X_i takes action s_i . Thus, the probability of the set of actions that yield c_i is

$$\bar{P}_i [f_i^{-1}(c_i)] = \sum_{s_j \in f_i^{-1}(c_i)} \bar{p}_i(s_j). \quad (63)$$

Let $\{j_{i1}, \dots, j_{iN_i}\}$ denote a permutation of $\{1, \dots, N_i\}$ such that

$$\bar{w}_i(x_{ij_1}) \geq \dots \geq \bar{w}_i(x_{ij_{N_i}}) \quad (64)$$

and apply the Quiggin operators as defined by (55) to the converged probability measures \bar{P}_i , yielding the rank-dependent decision weights

$$\begin{aligned} \pi(x_{ij_{i1}}) &= \omega[\bar{P}_i [f_i^{-1}(x_{ij_{i1}})]] \\ \pi(x_{ij_{i2}}) &= \omega[\bar{P}_i [f_i^{-1}(x_{ij_{i1}})] + \bar{P}_i [f_i^{-1}(x_{ij_{i2}})]] - \omega[\bar{P}_i [f_i^{-1}(x_{ij_{i1}})]] \\ &\vdots \\ \pi(x_{ij_{ik}}) &= \omega[\bar{P} [f^{-1}(x_{ij_{i1}})] + \dots + \bar{P} [f^{-1}(x_{ij_{ik}})]] \\ &\quad - \omega[\bar{P} [f^{-1}(x_{ij_{i1}})] + \dots + \bar{P} [f^{-1}(x_{ij_{i(k-1)}})]] \\ &\vdots \\ \pi(x_{ij_{iN_i}}) &= \omega(1) - \omega[\bar{P}_i [f_i^{-1}(x_{ij_{i1}})] + \dots + \bar{P}_i [f_i^{-1}(x_{ij_{i(N_i-1)}})]] \end{aligned} \quad (65)$$

and the rank-dependent utility of \tilde{f}_i is

$$RDU[\bar{w}_i(\tilde{f}_i)] = \sum_k \bar{w}_i(x_{ij_k}) \pi(x_{ij_k}). \quad (66)$$

The presence in a social influence network of agents whose choices reflect RDU is thus picked up in the transition matrices that, from Section 3.4, provide correlating signals for unconditional play. We reflect on the significance of this in the concluding section.

5 Concluding Discussion

We argued in Section 3 that CGT analysis can represent strategic mindshaping, and that such analysis can provide insight into pre-play processes that furnish common knowledge among players, which can in turn be used as a basis for identifying correlated equilibria following Aumann (1987). In Section 4 we argued that if some players have RDU preferences, these can be reflected in the conditional game, and their influence is picked up in the transition matrices that constitute the shared signal for the unconditional stage of analysis.

The intended significance of the first argument is straightforward. Mindshaping processes make people, at least within the shared cultural spaces where complex coordination is observed, sufficiently predictable to one another to support such coordination. Therefore, that is the process

we should hypothesize as furnishing signals for correlated equilibrium *if* we think that Aumann's purely theoretical argument is relevant to explaining how actual people find equilibrium in games where institutional rules are insufficiently precise and binding to select an equilibrium for them. Mindshaping has a strategic aspect, and therefore resources from game theory should be used, if possible, to model it. That is the purpose of CGT.

The significance of the second argument is more speculative - but, to the extent that our speculation is thought to be persuasive, more interesting and important.

Deciding which mathematical theory to use to model empirical events involves interpretive decisions. We refer here not to identification through measurement protocols, but to deeper decisions about how to map abstract mathematical elements onto phenomena being measured. Confronted by the fact that (typically) majorities of subjects in economic experiments choose in ways that indicate that personal subjective probabilities influence their utility functions, an experimenter might conclude that she should not try to apply theory that relies on the Harsanyi Doctrine for purpose of analyzing behavior. Aumann concludes his 1987 paper with an insightful discussion of the implications of such a decision. He points out that if agents have different prior probabilities even after taking account of the expectation that all are Bayesians, then one can still define a *subjective correlated equilibrium* in which agents recognize that they each draw varying priors from a common distribution. However, the result of applying this concept to equilibrium selection is disappointingly weak. Aumann provides examples of simple games in which sets of subjective correlated equilibria include both optimal and sub-optimal strategy vectors. Another path, Aumann notes, is followed by the rationalizability literature Bernheim (1984); Pearce (1984). But rationalizability is the weakest of equilibrium refinements.

Our argument of Section 4 offers an alternative strategy. This is to interpret such observed 'personal probabilities' as are not simply false perceptions or miscalculations as reflecting uncertainty about the background strategic conditions of a game. The potential rationalisations for subjective decision weights discussed by Harrison and Ross (2018), and mentioned in Section 4, are of this type. Those are just the sorts of preference equivocations that CGT is designed to resolve. Since they are reflected in the transition matrices for unconditional games that the CGT pre-play generates as common knowledge, one might then argue that the game theory modeler should not re-introduce them again into the utilities at play in the unconditional analysis; she should rather model the players as already taking them into account, and build her game on the maintained assumption that players are Savage expected utility maximizers who have accurate estimates of expected frequencies of observed choices that violate EUT in specified ways. Then she can apply the Harsanyi Doctrine as per Aumann's clarification of it. That is, players will update their estimates differently as they each obtain private information from the course of unconditional play, and the posterior probabilities they apply will consequently differ. But the basis for correlated equilibrium play is modeled as intact.

It is worth repeating our earlier point that the audience to whom this suggestion is directed is not the typical first-order experimenter. She will want to directly estimate her subjects' risk preferences using procedures of the kind surveyed in Harrison and Rustrom (2008), and she will typically find that many of the subjects are best characterized by RDU. If her experiment involves identifying equilibria in an experimental game, she will likely prefer to use an error-tolerant solution concept such as quantal response equilibrium, and treat the equilibrium selection process as a black box.

On the other hand, some experiments are directly about equilibrium learning by subjects. If the situation faced by these subjects is novel to them, an experimenter might have them participate

in some initial learning rounds. Typically such rounds are unincentivized, since it is assumed that their point is simply to help subjects understand the rules of the game, and recognize moves they might make to test hypotheses about other players' strategies. But this assumes that subjects are not inspired by the novelty of the setting to engage in mindshaping by conditioning their choices on other players' actions and adjusting preferences over outcomes accordingly. A way to check this assumption would be to separately incentivize learning rounds and estimate a CGT model of the resulting data from those rounds. If evidence of mindshaping emerged, then our arguments here would generate an empirical prediction: if the game has multiple equilibria, the frequency of identifying relatively efficient equilibria by the players should increase.

Our primary point, however, is for theorists. Controversies about how to understand the welfare implications of patterns of choice that reflect subjective decision weights stems from uncertainty about whether such behavior should be taken as revealing preferences that should be respected by a non-paternalist policy-maker or advisor, or as errors in understanding probability that the advisor should helpfully correct before pronouncing a recommendation or designing a nudge. This is an issue of philosophical interpretation, with normative implications. It is not technically resolvable, because it stems ultimately from the fact that if the externalist position in cognitive science is correct, then there are no psychological facts about *latent* preferences and beliefs that could be discovered to empirically settle the issue. However, our practical suggestion about how to think about the issue offers a way to avoid simply living with muddled intuitions while philosophical decision theorists debate indefinitely about what 'rationality' means. When people are engaged in unresolved mindshaping some of their preferences remain indeterminate. But where choice patterns have observably settled down, mindshaping can be regarded as having locally done its work. Then decision weights might most reasonably be treated as preferences just like any others - stable results of mindshaping processes is just what preferences *are* in the first place. Here our ability, using CGT, to rigorously frame mindshaping as *strategic* matters normatively: advisors should not, in general, set out to defeat the strategies of their clients. But against the backdrop of such stability, choices that *do* result from errors can be identified and, when the advisor's confidence meets scientific standards of evidence, corrected.⁹

References

- Aumann, R.J., 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1, 67–96.
- Aumann, R.J., 1987. Correlated equilibrium as an expression of bayesian rationality. *Econometrica* 55, 1–18.
- Bernheim, B.D., 1984. Rationalizable strategic behavior. *Econometrica* 52, 1007–1028.
- Bonabeau, E., Dorigo, M., Theraulaz, G., 1999. *Swarm Intelligence*. Oxford University Press.
- Bourgeois-Gironde, S., 2020. *The Mind Under the Axioms*. Academic Press.

⁹What is described here is a special application, to the interpretation of subjective decision weights, of the general strategy for economic interpretation of behavior that Harrison and Ross call the "Quantitative Intentional Stance"; see Harrison and Ross (2023) for the general account.

-
- Bradley, R., 2017. *Decision Theory with a Human Face*. Cambridge University Press.
- Buchak, L., 2013. *Risk and Rationality*. Oxford University Press.
- Camerer, C., 2003. *Behavioral Game Theory*. Princeton University Press.
- Chittka, L., 2022. *The Mind of a Bee*. Princeton University Press.
- Clark, A., 1997. *Being There*. MIT Press.
- Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J., 1999. *Probabilistic Networks and Expert Systems*. Springer Verlag.
- Dennett, D., 1987. *The Intentional Stance*. MIT Press.
- Dennett, D., 2017. *From Bacteria to Bach and Back*. Allen Lane.
- Doob, J.L., 1953. *Stochastic Processes*. John Wiley & Sons.
- Frith, C., Wolpert, D. (Eds.), 2004. *The Neuroscience of Social Interaction*. Oxford University Press.
- Fudenberg, D., Levine, D., 1998. *The Theory of Learning in Games*. MIT Press.
- Harrison, G., Martinez-Correa, J., Swarthout, J.T., 2015. Reduction of compound lotteries with objective probabilities. *Journal of behavior and organization* 119, 32–55.
- 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., Ross, D., 2018. Varieties of paternalism and the heterogeneity of utility structures. *Journal of Economic Methodology* 25, 42–67.
- Harrison, G., Ross, D., 2023. Behavioral welfare economics and the quantitative intentional stance, in: Harrison, G.W., Ross, D. (Eds.), *Models Of Risk Preferences: Descriptive And Normative Challenges*. Emerald. Forthcoming.
- Harrison, G., Ruström, E.E., 2008. Risk aversion in the laboratory, in: Cox, J., Harrison, G. (Eds.), *Risk Aversion in Experiments*, pp. 41–196.
- Harsanyi, J., 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge University Press.
- Hutto, D., 2008. *Folk Psychological Narratives*. MIT Press.
- Jensen, F.V., 2001. *Bayesian Networks and Decision Graphs*. Springer Verlag.
- Jordan, M.I. (Ed.), 2001. *Learning in Graphical Models*. MIT Press.

- Kreps, D., Wilson, R., 1982. Sequential equilibria. *Econometrica* 50, 863–984.
- Lauritzen, S.L., 1996. *Graphical Models*. Oxford University Press.
- Lichtenstein, S., Slovic, P. (Eds.), 2006. *The Construction of Preference*. Cambridge University Press.
- Malle, B., Moses, L., Baldwin, D. (Eds.), 2003. *Intensions and Intentionality*. MIT Press.
- Nichols, S., Stich, S. (Eds.), 2002. *Mindreading*. Oxford University Press.
- Pearce, D., 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52, 1029–1050.
- Pearl, J., 1988. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.
- Pearl, J., 2009. *Causality: Models, Reasoning, and Inference*. second ed., Cambridge University Press, Cambridge, UK.
- 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.
- Quiggin, J., 1993. *Generalized Expected Utility Theory*. Kluwer.
- Ross, D., 2004. Meta-linguistic signalling for coordination amongst social agents. *Language Sciences* 52, 621–642.
- Ross, D., 2006. The economics and evolution of selves. *Journal of Cognitive Systems Research* 7, 246–258.
- Ross, D., 2023. Neo-Samuelsonian methodology, normative economics, and the quantitative intentional stance, in: Caldwell, B., Davis, J., Mäki, U., Sent, E.M. (Eds.), *Methodology And History Of Economics*. Routledge, pp. 90–118.
- Ross, D., Stirling, W.C., 2021. Economics, social neuroscience, and mindshaping, in: Harbecke, J., Herrmann-Pillath, C. (Eds.), *Social Neuroeconomics*. Routledge, pp. 174–201.
- Savage, L.J., 1954. *The Foundations of Statistics*. John Wiley & Sons.
- Sprites, P., Glymour, C., Scheines, R., 2000. *Causation, Prediction, and Search*. second ed., MIT 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.
- Zawidzki, T.W., 2013. *Mindshaping: A New Framework for Understanding Human Social Cognition*. MIT Press.