Language-based Games

Adam Bjorndahl, Joseph Y. Halpern, Rafael Pass
Cornell University

Abstract

We introduce language-based games, in which utility is defined over descriptions in a given language. By choosing the right language, we can capture psychological games [9] and reference-dependent preference [13]. Of special interest are languages that can express only coarse beliefs (e.g., the probability of an event is "high" or "low", rather than "the probability is .628"): by assuming that a player's preferences depend only on what is true in a coarse language, we can resolve a number of well-known paradoxes in the literature, including the Allais paradox. Despite the expressive power of this approach, we show that it can describe games in a simple, natural way. Nash equilibrium and rationalizability are generalized to this setting; Nash equilibrium is shown not to exist in general, while the existence of rationalizable strategies is proved under mild conditions on the language.

1 Introduction

In a classical, normal-form game, an *outcome* is a tuple of strategies, one for each player, and players' preferences are formalized by utility functions defined on the set of all such outcomes. This framework thereby hard-codes a single conception of how players represent the world insofar as their preferences are concerned.

The motivating idea of the present work is to relax this rigidity in a systematic way by using *language* as the foundation of preference. Roughly speaking, we assume that what the players care about is captured by some *underlying language*, with utility defined on *descriptions* in that language. Classical game theory can be viewed as the special case where the underlying language can talk only about outcomes. In general, however, the language can be as rich or poor as desired.

In the colloquial sense of the word, the role of "language" in decision making and preference formation can hardly be overstated. It is well known, for example, that presenting alternative medical treatments in terms of survival rates versus mortality rates can produce a marked difference in how those treatments are evaluated, even by experienced physicians [16]. More generally, one of the core insights of *prospect theory* [REF]—that subjective value depends not (only) on facts about the world but on how those facts are *presented* (as gains or losses, dominated or undominated options, etc.)—can be viewed as a kind of language-sensitivity. We celebrate 10th and 100th anniversaries specially, and make a big deal when the Dow Jones Industrial Average crosses a multiple of 1,000, all because we happen to work in a base 10 number system (i.e., our language puts special emphasis on multiples of 10 that would be absent, for example, in a hexadecimal system). Furthermore, we often assess likelihoods using words like "probable", "unlikely", or "negligible", rather than numeric representations, and when numbers are used, we tend to round them [15]. Much of the motivation and conceptual appeal of our approach stems from observations like these: defining preferences in terms of language provides a direct avenue for formalizing such intuitions about how people think.

Of special interest is the general phenomenon of *coarseness* or *categoricity*. Theories of rational decision making are often couched in the formalism of continuous mathematics, but the world is not always a continuous place, at least as far as preferences are concerned. Consumers tend to ignore, for example, the difference in price between $3.98 and $3.99, but take seriously (or even exaggerate) the difference between $3.99 and $4.00 [REF] (cf. Example 3.1). Similarly, although degrees of belief are often formalized using probability measures, a coarser representation can be more appropriate for reasoning about human choice and inference (see [**?**], [18], [15]). We show, for instance, that the Allais paradox [1] can be resolved simply and intuitively when belief is represented discretely, rather than on a continuum (Example 3.2).

Coarseness in the underlying language—cases where there are fewer descriptions than there are actual differences to describe—provides a natural and powerful way of capturing such phenomena, offering insight into a variety of puzzles and paradoxes of human decision making. Moreover, it allows for a unified analysis of coarseness as it pertains both to preferences and to beliefs, traditionally distinct domains of decision making. This is accomplished using languages expressive enough to talk about beliefs, a technique that is of interest in its own right.

Classically, beliefs are relevant to decision making insofar as they determine *expected utility*. But beliefs can also themselves be considered as objects of preference: one might wish to model players who feel guilt, wish to surprise their opponents, or are motivated by a desire to live up to what is expected of them. *Psychological game theory*, beginning with the work of Geanakoplos, Pearce, and Stachetti [9] and expanded by Battigalli and Duwfenberg [4], is an enrichment of the classical setting meant to capture such preferences and motivations. In a similar vein, the notion of *reference-dependent preferences* developed by Köszegi and Rabin [13], building on prospect theory, formalizes phenomena such as loss-aversion by augmenting players' preferences with an additional sense of gain or loss derived by comparing the actual outcome to what was expected.

With the appropriate choice of language, our approach subsumes these: an underlying language that includes beliefs allows us to capture psychological games, while a language that distinguishes expected from actual outcomes allows us to represent reference-dependent preferences. Moreover, in each of these frameworks, modeling coarse beliefs provides insight and opportunities lacking in the continuous setting. Much of this paper is an elaboration and justification of this point.

As a preliminary illustration of some of these ideas, consider the following simple example.

Example 1.1: *A surprise proposal.* Alice and Bob have been dating for a while now, and Bob has decided that the time is right to pop the big question. Though he is not one for fancy proposals, he does want it to be a surprise. In fact, if Alice expects the proposal, Bob would prefer to postpone it entirely until such time as it might be a surprise. Otherwise, if Alice is not expecting it, Bob's preference is to take the opportunity.

We might summarize this scenario by the following table of payoffs for Bob:

	p	$\neg p$
$B_A\,p$	0	1
$\neg B_A\,p$	1	0

Table 1: The surprise proposal.

In this table, we denote Bob's two strategies, proposing and not proposing, by p and $\neg p$, respectively, and use $B_A p$ (respectively, $\neg B_A p$) to denote that Alice is expecting (respectively, not expecting) the proposal. Of course, whether or not Alice expects a proposal may be more than a binary affair: she may, for example, consider a proposal unlikely, somewhat likely, very likely, or certain. But as we have discussed, there is good reason to think that an accurate model of her expectations involves only a small number k of distinct "levels" of belief, rather than a continuum. Table 1, for simplicity, assumes that $k = 2$, though this is easily generalized to larger values.

Note that although Alice does not have a choice to make (formally, her strategy set is a singleton), she does have beliefs about which strategy Bob will choose. To represent Bob's preference for a surprise proposal, we must incorporate Alice's beliefs about Bob's choice of strategy into Bob's utility function. In psychological game theory, this is accomplished by letting $\alpha \in [0, 1]$ be the probability that Alice assigns to Bob proposing, and defining Bob's utility function u_B in some simple way so that it is decreasing in α if Bob chooses to propose, and increasing in α otherwise,[1] as for instance in the following:

$$u_B(x, \alpha) = \begin{cases} 1 - \alpha & \text{if } x = p \\ \alpha & \text{if } x = \neg p. \end{cases}$$

The function u_B agrees with Table 1 at its extreme points if we identify $B_A p$ with $\alpha = 1$ and $\neg B_A p$ with $\alpha = 0$. Otherwise, for the continuum of other values that α may take between 0 and 1, u_B yields a linear combination of the corresponding extreme points. Thus, in a sense, u_B is a continuous approximation to a scenario that is essentially discrete.

By contrast, we view Table 1 as *defining* Bob's utility. To coax an actual *function* from this table, let the variable S denote a *situation*, which for the time being we can conceptualize as a collection of statements about the game; in this case, S includes whether or not Bob is proposing, and whether or not Alice believes he is proposing. We then define

$$u_B(S) = \begin{cases} 0 & \text{if } p \in S \text{ and } B_A\, p \in S \\ 1 & \text{if } p \in S \text{ and } \neg B_A\, p \in S \\ 1 & \text{if } \neg p \in S \text{ and } B_A\, p \in S \\ 0 & \text{if } \neg p \in S \text{ and } \neg B_A\, p \in S. \end{cases}$$

In other words, Bob's utility is a function not merely of the outcome of the game (p or $\neg p$), but of a more general object we call a "situation"; his utility in a given situation S depends on his own actions combined with Alice's beliefs in exactly the manner prescribed by Table 1. As noted above, we may very well wish to refine our representation of Alice's state of surprise using more than two categories of likelihood; we could even allow a representation that permits continuous probabilities, as has been done in the literature. We spell out these straightforward generalizations in Example 3.5. ∎

The central concept we develop in this paper is that of a *language-based game*, where utility is defined not on outcomes but on *situations*. As noted, a situation can be conceptualized as a collection of statements about the game; intuitively, each statement is a description of something that might be relevant to a player's preferences, such as whether or not Alice believes that Bob will play a certain strategy. Of course, this notion crucially depends on just what counts as an admissible description. The set of all admissible descriptions—what we refer to as the *underlying language* of the game—is a key component of our model. Since utility is defined on situations, and situations are sets of descriptions taken from the underlying language, a player's preferences can depend, in principle, on anything expressible in this language, but nothing more. Succintly: players can prefer one state of the world to another if and only if they can *describe* the difference between the two in the underlying language.

[1] Technically, Geanakoplos et al. [9] allow Bob's utility to be a function of only his own beliefs; this is generalized by Battigalli and Duwfenberg [4] in the context of extensive-form games, but their approach is applicable to normal-form games as well.

From a technical standpoint, this paper makes three major contributions. First, we define a generalization of classical game theory and demonstrate its versatility in modeling a wide variety of strategic scenarios, focusing in particular on psychological and reference-dependent effects. Second, we provide a formal representation of coarse beliefs in a game-theoretic context. This exposes an important insight: a discrete representation of belief, often conceptually and technically easier to work with than its continuous counterpart, is sufficient to capture psychological phenomena that have heretofore been modeled only in a continuous framework. Moreover, as we show by example, utilities defined over coarse beliefs provide a natural way of capturing some otherwise puzzling behavior. Third, we provide novel equilibrium analyses for a broad class of language-based games that do not depend on continuity assumptions as do those of, for example, Geanakoplos et al. [9]. In particular, our main theorem demonstrates that if the underlying language satisfies certain natural "compactness" assumptions, then every game over this language admits rationalizable strategies. By contrast, even under such compactness assumptions, not every game admits a Nash equilibrium (see Example 3.4).

The rest of the paper is organized as follows. In Section 6.1, we develop the basic apparatus needed to describe our approach. Section 3 presents a collection of examples intended to guide intuition and showcase the system. In Section 4, we show that there is a natural route by which solution concepts such as Nash equilibrium and rationalizability can be defined in our setting, and we address the question of existence. Section 5 is an in-depth analysis of an example studied by Köszegi and Rabin [13], interpreted as a langugage-based game. Appendix A collects the proofs that are omitted from the main body.

2 Foundations

2.1 Game forms and intuition

Much of the familiar apparatus of classical game theory is left untouched. A **game form** is a tuple $\Gamma = (N, (\Sigma_i)_{i \in N})$ where N is a finite set of *players*, which for convenience we take to be the set $\{1, \dots, n\}$, and Σ_i is the set of *strategies available to player i*. Following standard notation, we set

$$\Sigma := \prod_{i \in N} \Sigma_i \quad \text{and} \quad \Sigma_{-i} := \prod_{j \neq i} \Sigma_j.$$

Elements of Σ are called *outcomes* or *strategy profiles*; given $\sigma \in \Sigma$, we denote by σ_i the ith component of the tuple σ, and by σ_{-i} the element of Σ_{-i} consisting of all but the ith component of σ.

Note that a game form does not come equipped with utility functions specifying the preferences of players over outcomes Σ. The utility functions that we employ are defined on situations, which in turn are determined by the underlying language, so, before defining utility, we must first formalize these notions.

Informally, a *situation* is an exhaustive characterization of a given state of affairs using descriptions drawn from the underlying language. Assuming for the moment that we have access to a fixed "language", we might imagine a situation as being generated by simply listing all statements from that language that happen to be true of the world. Even at this intuitive level, it should be evident that the informational content of a situation is completely dependent on the expressiveness of the language. If, for example, the underlying language consists of exactly two descriptions, "It's raining" and "It's not raining", then there are only two situations:

$$\{\text{“It's raining”}\} \quad \text{and} \quad \{\text{“It's not raining”}\}.$$

More formally, a situation S is a set of formulas drawn from a larger pool of well-formed formulas, the underlying language. We require that S include as many formulas as possible without being contradictory; this is made precise below.

The present formulation, informal though it is, is sufficient to allow us to capture a claim made in the introduction: any classical game can be recovered in our framework with the appropriate choice of underlying language. Specifically, let the underlying language be Σ, the set of all strategy profiles. Situations, in this case, are simply singleton subsets of Σ, as any larger set would contain distinct and thus intuitively contradictory descriptions of the outcome of the game. The set of situations can thus be identified with the set of outcomes, so a utility function defined on outcomes is readily identified with one defined on situations.

In this instance the underlying language, consisting solely of atomic, mutually incompatible formulas, is essentially structureless; one might wonder why call it a "language" at all, rather than merely a "set". Although, in principle, there are no restrictions on the kinds of objects we might consider as languages, it can be very useful to focus on those with some internal structure. This structure has two aspects: syntactic and semantic.

2.2 Syntax, semantics, and situations

A formal language is typically generated from a set of atomic formulas using some rules. For example, given a set Φ of *primitive propositions*, let $\mathcal{L}(\Phi)$ denote the language generated by starting with the formulas in Φ, and then closing off under conjunction ($\wedge$) and negation ($\neg$). (We can define $\vee$ and $\rightarrow$ from $\neg$ and $\wedge$ as usual.) $\mathcal{L}(\Phi)$ is a language for reasoning about Boolean combinations of the propositions in Φ. This is easily specialized to a game-theoretic setting. Given a game form $\Gamma = (N, (\Sigma_i)_{i \in N})$, let

$$\Phi_\Gamma = \{play_i(\sigma_i) \,:\, i \in N,\, \sigma_i \in \Sigma_i\},$$

where we read $play_i(\sigma_i)$ as "player i is playing strategy σ_i". Then $\mathcal{L}(\Phi_\Gamma)$ is a language appropriate for reasoning about the strategies chosen by the players in Γ. We sometimes write $play(\sigma)$ as an abbreviation for $play_1(\sigma_1) \wedge \cdots \wedge play_n(\sigma_n)$.

Semantics provides a notion of truth. Recall that the semantics of classical propositional logic is given by *valuations* $v : \Phi \to \{\mathsf{true}, \mathsf{false}\}$. Valuations are extended to all formulas via the familiar truth tables for the logical connectives. Each valuation v thereby generates a *model*, determining the truth values of every formula in $\mathcal{L}(\Phi)$. In the case of the language $\mathcal{L}(\Phi_\Gamma)$, we restrict this class of models to those corresponding to an outcome $\sigma \in \Sigma$; that is, we consider only valuations v_σ defined by

$$v_\sigma(play_i(\sigma'_i)) = \mathsf{true} \text{ iff } \sigma'_i = \sigma_i,$$

so that, intuitively, each player chooses exactly one strategy. Denote this restricted class of models by $\mathcal{M}(\Gamma)$.

Given a language $\mathcal{L}$ and a class $\mathcal{M}$ of models for $\mathcal{L}$, a set F of formulas in $\mathcal{L}$ is said to be **satisfiable in $\mathcal{M}$** there is some model in $\mathcal{M}$ in which every formula of F is true. An **$(\mathcal{L}, \mathcal{M})$-situation** is then defined to be a maximal set of formulas in $\mathcal{L}$ that is satisfiable in $\mathcal{M}$; that is, a satisfiable set with no proper superset that is also satisfiable. In the game-theoretic setting, as we have seen, each model in $\mathcal{M}(\Gamma)$ makes exactly one of the formulas $play_i(\sigma_i)$ true for each player i, so an $(\mathcal{L}(\Phi_\Gamma), \mathcal{M}(\Gamma))$-situation can be identified with a strategy profile. We denote by $\mathcal{S}(\mathcal{L}, \mathcal{M})$ the set of $(\mathcal{L}, \mathcal{M})$-situations. A game form Γ is extended to an **$(\mathcal{L}, \mathcal{M})$-game** by adding utility functions $u_i : \mathcal{S}(\mathcal{L}, \mathcal{M}) \to \mathbb{R}$, one for each player $i \in N$. $\mathcal{L}$ is called the **underlying language (of the game)**. We omit $\mathcal{L}$ and/or $\mathcal{M}$ when talking about situations or games whenever it is safe to do so.

In an $(\mathcal{L}(\Phi_\Gamma), \mathcal{M}(\Gamma))$-game, as observed above, the players' utility functions are essentially defined on Σ, so an $(\mathcal{L}(\Phi_\Gamma), \mathcal{M}(\Gamma))$-game is really just a standard normal-form game based on Γ. As we saw in Section 2.1, this class of games can also be represented with the completely structureless language Σ. This may well be sufficient for certain purposes, especially in cases where all we care about are two or three formulas. However, the structure of an underlying language $\mathcal{L}$ can be a powerful tool for studying the corresponding class of $\mathcal{L}$-games; in particular, a highly structured underlying language makes it easier to analyze the much broader class of psychological games.

A psychological game is an extension of a standard normal-form game except that players' preferences can depend not only on what strategies are played, but also on what beliefs are held. While $\mathcal{L}(\Phi_\Gamma)$ is appropriate for reasoning about strategies, it cannot express anything about beliefs. For this, we use a standard modal logic of belief [8].

Fix a game form $\Gamma = (N, (\Sigma_i)_{i \in N})$. Let $\mathcal{L}_B(\Phi_\Gamma)$ be the language obtained by starting with the formulas in Φ, then closing off under $\wedge$, $\neg$, and the unary operators B_i for $i = 1, \ldots, n$, so that if φ is a formula, so is $B_i\varphi$. We read $B_i\varphi$ as "player i believes φ". We also make use of the abbreviation $\widehat{B}_i$ for $\neg B_i \neg$, and read $\widehat{B}_i\varphi$ as "player i considers φ to be possible". Intuitively, $\mathcal{L}_B(\Phi_\Gamma)$ is a language for reasoning about the beliefs of the players and the strategies being played.

We give semantics to $\mathcal{L}_B(\Phi_\Gamma)$ using a standard modal logic construction [10]; for many applications of interest, understanding the (completely standard, although somewhat technical) details is not necessary. Example 1.1 was ultimately analyzed as an $\mathcal{L}_B(\Phi_\Gamma)$-game, despite the fact that we had not even defined the syntax of this language at the time, let alone its semantics. Section 3 provides more illustrations of this point.

A **Γ-structure** is a tuple $M = (\Omega, (s_i)_{i \in N}, (\mathcal{PR}_i)_{i \in N})$ satisfying the following conditions:

(P1) Ω is a nonempty measurable space;

(P2) $\mathcal{PR}_i : \Omega \to \Delta(\Omega)$ is measurable;

(P3) $\{\omega' : \mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)\}$ is measurable and $\mathcal{PR}_i(\omega)(\{\omega' : \mathcal{PR}_i(\omega') = \mathcal{PR}_i(\omega)\}) = 1$;

(P4) $s_i : \Omega \to \Sigma_i$ is such that $\{\omega' : s_i(\omega') = s_i(\omega)\}$ is measurable and $\mathcal{PR}_i(\omega)(\{\omega' : s_i(\omega') = s_i(\omega)\}) = 1$.

The set Ω is called the **state space**; $\Delta(\Omega)$ denotes the measurable space of all probability measures on Ω equipped with the σ-algebra generated by all sets of the form $\{\mu : \mu(E) = 1\}$, for $E \subseteq \Omega$ measurable. Conditions (P1) and (P2) set the stage to represent player i's beliefs at state $\omega \in \Omega$ using the probability measure $\mathcal{PR}_i(\omega)$ over the state space itself. Condition (P3) says essentially that players are sure of their own beliefs. The functions s_i are called the **strategy functions**, assigning to each state a strategy that we think of as what player i is playing at that state. Condition (P4) thus asserts that each player is sure of his own strategy. These assumptions are standard when representing belief in a game-theoretic setting [2].

The language $\mathcal{L}_B(\Phi_\Gamma)$ can be interpreted in any Γ-structure M via the strategy functions, which induce a valuation $[\![\cdot]\!]_M : \mathcal{L}_B(\Phi_\Gamma) \to 2^\Omega$ defined recursively by:

$$\begin{array}{lcl}
[\![play_i(\sigma_i)]\!]_M & := & \{\omega \in \Omega \, : \, s_i(\omega) = \sigma_i\} \\
[\![\neg\varphi]\!]_M & := & \Omega \setminus [\![\varphi]\!]_M \\
[\![\varphi \wedge \psi]\!]_M & := & [\![\varphi]\!]_M \cap [\![\psi]\!]_M \\
[\![B_i\varphi]\!]_M & := & \{\omega \in \Omega \, : \, \mathcal{PR}_i(\omega)([\![\varphi]\!]_M) = 1\}.
\end{array}$$

Thus, the Boolean connectives are interpreted classically, and $B_i\varphi$ holds at state ω just in case φ corresponds to a probability 1 event according to the measure

$\mathcal{PR}_i(\omega)$. It is easy to show (by induction on the structure of φ) that each set $[\![\varphi]\!]_M$ is measurable, so in particular the definition of $[\![B_i\varphi]\!]_M$ makes sense.

Let $\mathcal{M}_B(\Gamma)$ consist of all pairs of the form (M, ω), where $M = (\Omega, \vec{s}, \vec{\mathcal{PR}})$ is a Γ-structure and $\omega \in \Omega$. Given $\varphi \in \mathcal{L}_B(\Phi_\Gamma)$, we sometimes write $(M, \omega) \models \varphi$ instead of $\omega \in [\![\varphi]\!]_M$, and say that ω **satisfies** φ or φ is **true at** ω; we write $M \models \varphi$ and say that φ is **valid in** M if $[\![\varphi]\!]_M = \Omega$. Given $F \subseteq \mathcal{L}_B(\Phi_\Gamma)$, we write $(M, \omega) \models F$ if for all $\varphi \in F$, $(M, \omega) \models \varphi$.

It is not hard to see that when there is more than one player, $\mathcal{S}(\mathcal{L}_B(\Phi_\Gamma), \mathcal{M}_B(\Gamma))$ is infinite. A utility function $u_i : \mathcal{S}(\mathcal{L}_B(\Phi_\Gamma), \mathcal{M}_B(\Gamma)) \to \mathbb{R}$ can therefore be quite complicated. We will frequently be interested in representing preferences that are much simpler. For instance, though the surprise proposal scenario presented in Example 1.1 can be viewed as an $(\mathcal{L}_B(\Phi_\Gamma), \mathcal{M}_B(\Gamma))$-game, Bob's utility u_B does not depend on any situation as a whole, but rather is determined a small set of formulas. This motivates the following general definition, identifying a particularly easy to understand and well-behaved subclass of games.

Fix a language $\mathcal{L}$ and a class of models $\mathcal{M}$ for $\mathcal{L}$. A function $u : \mathcal{S}(\mathcal{L}, \mathcal{M}) \to \mathbb{R}$ is called **finitely specified** if there is a finite[2] set of formulas $F \subset \mathcal{L}$ and a function $f : F \to \mathbb{R}$ such that every situation $S \in \mathcal{S}(\mathcal{L}, \mathcal{M})$ contains exactly one formula from F, and whenever $\varphi \in S \cap F$, $u(S) = f(\varphi)$. In other words, the value of u depends only on the formulas in F. Thus, u is finitely specified if and only if it can be written in the form

$$u(S) = \begin{cases} a_1 & \text{if } \varphi_1 \in S \\ \vdots & \vdots \\ a_k & \text{if } \varphi_k \in S, \end{cases}$$

for some $a_1, \ldots, a_k \in \mathbb{R}$ and $\varphi_1, \ldots, \varphi_k \in \mathcal{L}$.

A language-based game is called finitely specified if each player's utility function is. Many games of interest are finitely specified. In a finitely specified game, we can think of a player's utility as being a function of the finite set F; indeed, we can think of the underlying language as being the structureless "language" F rather than $\mathcal{L}$.

3 Examples

We now provide a range of examples to exhibit both the simplicity and the expressive power of the language-based approach. Since we focus on the language $\mathcal{L}_B(\Phi_\Gamma)$ and the corresponding class of models $\mathcal{M}_B(\Gamma)$, we write $\mathcal{S}$ to abbreviate $\mathcal{S}(\mathcal{L}_B(\Phi_\Gamma), \mathcal{M}_B(\Gamma))$.

[2] If $(\mathcal{L}, \mathcal{M})$ is *compact* (see Section 4.3) then this finiteness condition on F is redundant. In particular, this holds for $(\mathcal{L}_B(\Phi_\Gamma), \mathcal{M}_B(\Gamma))$.

For each $S \in \mathcal{S}$ and each $i \in N$, note that there is a unique $\sigma_i \in \Sigma_i$ such that $play_i(\sigma_i) \in S$; we can think of σ_i as the strategy that player i is playing in the situation S. As such, when describing the utility of a situation, it is often useful to extract this strategy; therefore, we define $\rho_i : \mathcal{S} \to \Sigma_i$ implicitly by the requirement $play_i(\rho_i(S)) \in S$. It is easy to check that ρ_i is well-defined.

Example 3.1: *Preparing for a roadtrip.* Alice has two tasks to accomplish before embarking on a cross-country roadtrip: she needs to buy a suitcase, and she needs to buy a car.

Here we sketch a simple decision-theoretic scenario in a language-based framework to illustrate the power of coarseness. In particular, we choose the underlying language in such a way as to capture two well-known "irrationalities" of consumers. First, consumers often evaluate prices in a discontinuous way, behaving, for instance, as if the difference between \$299 and \$300 is more substantive than the difference between \$300 and \$301. Second, consumers who are willing to put themselves out (for example, drive an extra 5 kilometers) to save \$50 on a \$300 purchase are often not willing to make the same sacrifice for the same savings on a \$20,000 purchase (see [**?**]).

Both of the irrationalities described above can be captured by assuming a certain kind of coarseness, specifically, that the language over which Alice forms preferences does not describe prices with infinite precision. Consider a language with primitive propositions of the form p_c, roughly interpreted as "the price is about c", equipped with semantics mapping these propositions to certain intervals of the real line. For instance, the language might consist of the propositions p_{280}, p_{290}, and p_{300}, interpreted respectively as the price ranges $[\$280, \$290)$, $[\$290, \$300)$, and $[\$300, \$310)$. Any utility function defined over such a language cannot distinguish prices that fall into the same interval. Thus, in the example above, Alice would consider the prices \$300 and \$301 to be effectively the same as far as her preferences are concerned. At the borderline between intervals, however, there is the potential for a "jump": we might reasonably model Alice as prefering a situation described by p_{290} rather than by p_{300}—in other words, prefering to spend "about \$290" rather than "about \$300".

In this context, a smart retailer would set the price of her product to be at the upper end of an interval; of course, this assumes that the retailer has an understanding of the language over which their consumer base forms preferences (and moreover that each consumer makes use of roughly the same language). While there is some intuitive reason to think that certain cultural facts (like the use of a base 10 number system) have an influence in this regard, clearly these are major assumptions. Extending the language-based framework so as to capture players who can reason about the language of their opponents is therefore a promising direction for future research.

The second irrationality discussed above can be captured by assuming that the underlying language is not only coarse, but is *coarser* at higher prices. For example, around the \$20,000 mark, we might suppose that the language contains only the propositions p_{19000}, p_{19500}, and p_{20000}, interpreted respectively as the price ranges $[\$19000, \$19500)$, $[\$19500, \$20000)$, and $[\$20000, \$20500)$. In this case, while Alice may prefer a price of $300 to a price of $350, she cannot prefer a price of $20,000 to a price of $20,050, because that difference cannot be described in the underlying language.

This kind of analysis has a certain intuitive appeal: the larger the number (or, more generally, the further removed something is, in space or time or abstraction), the more you "ballpark" it—the less precise your language is in describing it. Indeed, psychological experiments have demonstrated that Weber's law[3], traditionally applied to physical stimuli, also finds purchase in the realm of numerical perception: larger numbers are subjectively harder to discriminate from one another [17; 20]. This type of example, as well as the observation that it can be understood as an instance of Weber's law, is due to Thaler [**?**]. Our choice of underlying language represents the phenomenon simply, while exhibiting its explanatory power.

But is it appropriate to model Alice using the coarse language described above? Surely she has mastered the basics of the Arabic numeral system, and can perfectly well describe the difference between 300 and 301, or between 20,000 and 20,050. How can this be reconciled with the use of coarseness? Intuitively, we think of Alice as using *two* languages: there is the (typically quite rich) language used to describe the world in general, and the (typically much coarser) language over which utility is defined—the underlying language. In general, these two languages may be quite different. There may be, for example, English words or mathematical expressions that have no correlate in Alice's underlying language—there is a mathematical expression for each of the infinitely-many natural numbers, but this should not entail that every such distinction is faithfully rendered in the representation of the world that Alice uses when she makes decisions.[4]

The underlying language provides the model of the world that Alice uses when she has to make a decision and evaluate her preferences; it describes the features that are salient to her in a decision-making context. Of course, which features are salient may be context-dependent, and how one identifies the appropriate language (and semantics) for modeling a given scenario is an interesting and important question. In the analysis above, for example, where did the intervals come from? This is far from an idle concern, as the choice of boundaries can have nontrivial implications. In Alice's case, although she cannot prefer a price of $20,000 to a price of $20,050, she *can* prefer a price of $19,950 to a price of $20,000. Some might deny that this is a reasonable assumption, arguing

[3] Weber's law asserts that the minimum difference between two stimuli necessary for a subject to discriminate between them is proportional to the magnitude of the stimuli; thus, larger stimuli require larger differences between them to be perceived.

[4] Conversely, there may be descriptions in the underlying language that don't correspond to anything in, say, English. In a blind taste test, Bob might say that he prefers one type of ice cream to others, without being able to express verbally what it is about the taste experience that leads to the preference.

perhaps that the interval corresponding to p_{20000} should be centered on \$20,000 instead, or denying the existence of sharp boundaries altogether. But even if sharp boundaries are accommodated, the question of their origin is critical, if for no other reason than their obvious potential for exploitation.

One natural way of modeling the context-dependence of the underlying language is to assume that, while the syntax is fixed, the meanings of words can depend on context. For example, we might think of a typical consumer as reasoning about prices with a fixed collection of categories (e.g. "free", "cheap", "good deal", "fair price", "a bit much", "expensive", and "prohibitive"), but the mapping from these terms in the underlying language to real cost varies depending on the shopping context. When buying a car, "cheap" might include prices of several hundred or even several thousand dollars, but when dining out similar prices for the entrées could well be considered "prohibitive". Another familiar case that seems to typify this pattern is found in the end-of-year assignment of letter grades to students in a class. Once again the syntax is fixed (A+, A, A–, B+, etc.), but what exactly counts as an A+ is not. In many cases this is determined on the fly, with instructors essentially eyeballing boundaries between categories in such a way as to coincide with gaps in the raw scores, precisely so as to avoid sharp discontinuities in grades. Cases like these are especially interesting because they give some insight into the mechanics that govern the determination of the mapping from language to the world. Such a context-dependent semantics might also help explain the well-known *anchoring* effect [**?**]: the first price an agent is exposed to might tend to be classified as "reasonable" or "fair", with all other categories being determined relative to that initial calibration. Integrating this conception of a variable semantics into the language-based framework is clearly an important direction for future research. ∎

Example 3.2: *The Allais paradox [1].* Consider the two pairs of gambles described in Table 2. The first pair is a choice between (1a) \$1 million for sure, versus (1b) a .89 chance of \$1 million, a .1 chance of \$5 million, and a .01 chance of nothing. The second is a choice between (2a) a .89 chance of nothing and a .11 chance of \$1 million, versus (2b) a .9 chance of nothing and a .1 chance

Gamble 1a		Gamble 1b	
1	\$1 million	.89	\$1 million
		.1	\$5 million
		.01	\$0

Gamble 2a		Gamble 2b	
.89	\$0	.9	\$0
.11	\$1 million	.1	\$5 million

Table 2: The Allais paradox

of $5 million. The "paradox" arises from the fact that most people choose (1a) over (1b), and most people choose (2b) over (2a) [1], but these preferences are not simultaneously compatible with expected utility maximization.

Coarseness in the language of preference offers a simple and intuitive explanation of this phenomenon, and by essentially the same mechanism at play in Example 3.1. Let us assume that probability judgements such as "there is a .11 chance of getting $1 million" are represented in a language with only finitely-many "levels" of likelihood. In particular, suppose the language has only the descriptions "no chance", "slight chance", "unlikely", and their respective opposites, "certain", "near certain", and "likely", interpreted as in Table 3. Suppose further that for the purposes of utility, "expected" values are calculated using approximations obtained by identifying each "level" of likelihood with its midpoint; thus, a "slight chance" is approximated as a .025 chance, a "likely" event as a .9 probability, and so on.

True likelihood	Description	Approximation
1	certain	1
$[.95, 1)$	near certain	.975
$[.85, .95)$	likely	.9
$(.05, .15]$	unlikely	.1
$(0, .05]$	slight chance	.025
0	no chance	0

Table 3: Coarse likelihood approximations

Revisting the gambles associated with the Allais paradox, we see that the rounding errors introduced by coarseness change Alice's evaluation of the gambles significantly (Table 4). For one thing, probabilities of .89 and .9 are not distinguished at all (nor are .1 and .11), which immediately implies that (2b) is preferred to (2a), provided $u_A(\$5\text{ million}) > u_A(\$1\text{ million})$. On the other hand, likelihoods of 0 and .01 are not only distinguished by this language, but their difference is effectively exaggerated. Table 5 shows the result of substituting the approximations from Table 3 in for the descriptions of Table 4. We can

Gamble 1a		Gamble 1b	
certain	$1 million	likely	$1 million
		unlikely	$5 million
		slight chance	$0

Gamble 2a		Gamble 2b	
likely	$0	likely	$0
unlikely	$1 million	unlikely	$5 million

Table 4: The Allais pardox, coarsely described

Gamble 1a		Gamble 1b	
1	\$1 million	.9	\$1 million
		.1	\$5 million
		.025	\$0

Gamble 2a		Gamble 2b	
.9	\$0	.9	\$0
.1	\$1 million	.1	\$5 million

Table 5: The Allais paradox, coarsely approximated

calculate the revised utility of (1b) to be

$$.9 \cdot u_A(\$1 \text{ million}) + .1 \cdot u_A(\$5 \text{ million}) + .025 \cdot u_A(\$0),$$

and this quantity may well be less than $u_A(\$1 \text{ million})$, depending on the utility function u_A. For example, if

$$\begin{aligned} u_A(\$1 \text{ million}) &= 1 \\ u_A(\$5 \text{ million}) &= 3 \\ u_A(\$0) &= -10, \end{aligned}$$

then the utility of gamble (1b) evaluates to .95. In this case, Alice prefers (2b) to (2a) but also prefers (1a) to (1b).

Rubinstein [**?**] has offered a closely related analysis of the kind of reasoning that guides decision making in Allais-type environments. He suggests that agents may simplify some choice problems by "canceling" certain parameters that are judged to be sufficiently similar; for instance, the similarity between 0.1 and 0.11 might lead one to view gamble 2 as essentially a choice between \$1 million and \$5 million. Clearly this is very much in the same spirit as our analysis; indeed, Rubinstein goes on to observe that the same lottery may be subject to different similarity judgements depending on how it is presented. Thus, while he does not explicitly or formally invoke language as the object of preference (instead he develops a theory based on *similarity relations*), certainly much of the insight inherent in the use of language for the purpose of capturing coarseness effects is anticipated in his work.

It is worth taking a closer look at the particular type of coarseness we have employed here. With the exception of giving 0 its own category (an assumption that might reasonably be viewed as supported by psychological evidence [REFS]), the other boundaries appear rather arbitrary. Why should the probabilities 0.1 and 0.11 fall into the same category? A different partition could have separated them; indeed, it seems plausible that if Alice had instead been presented with gambles involving the probabilities 0.1, 0.101, 0.109, 0.11, and 0.111, then she may well have categorized 0.1 separately from 0.11. This is reminiscent of the case of the instructor assigning letter-grades considered in Example 3.1, and once again brings to attention the importance of understanding how categories are chosen—or, in our terms, how expressions in a given syntax are given semantics.

The fact that coarseness plays the pivotal role both in our analysis of the Allais paradox and also in understanding the behaviours discussed in Example 3.1 is notable because the domain of coarseness is quite different in these two cases. In Example 3.1, it is prices that are subject to conflation, whereas in this example it is degrees of belief. This example also emphasizes an important feature of language-based games: beliefs and preferences need not be as independent as they are in the standard framework. Classically, beliefs are relevant to decision-making only insofar as they determine expected utility; by contrast, in any language-based game where the underlying language can express beliefs, players can have preferences *about their own beliefs* (e.g., Alice can prefer to believe that her winning $1 million is guaranteed, rather than merely likely, even if in the second case she also believes there is a chance she will win $5 million). Thus, coarseness in the underlying language can manifest itself both in traditional objects of preference, like prices, but also in other areas relevant to decision-making, like beliefs. The use of language allows us to analyze such disparate examples as instances of the same general phenomenon. ∎

Example 3.3: *Playing the lottery.* Alice buys a lottery ticket, despite the fact that the purchase is, technically, an expected loss. Moreover, she buys only 1 ticket, not two, and not as many as she can afford.

One possible explanation for why people buy lottery tickets is that they are just *wrong* about the odds—they think the chance of winning such a large sum of money overcomes the cost of the ticket, rendering the transaction an expected gain. While it is reasonable that with such large and small numbers involved mistakes might be made, this is on the whole not a convincing account of the rationale behind playing the lottery. The flaw is very basic: if buying one ticket were evaluated as an expected gain, then buying two or ten thousand tickets should be viewed as even better.

In Example 3.2, we saw that coarseness in the underlying language can result in some differences of likelihood being collapsed, and others exaggerated. These types of rounding errors can also be employed to explain the allure of playing the lottery.

Consider a simple lottery in which each ticket costs $1 and provides a 0.00005% chance to win $1 million, so each ticket in fact yields an expected loss of 50¢. But suppose also that the lowest non-zero level of likelihood expressible in Alice's language of preference, the "slight chance" description, subsumes anything up to a 0.01% probability, and is evaluated for the purposes of utility calculations as a 0.005% probability. In this case, the expected value of purchasing a lottery ticket, which costs $1 and provides a "slight chance" of winning $1 million, jumps up to a gain of $49. Perhaps more striking, the expected value of purchasing *two* lottery tickets, which costs $2 but still provides, as far as the language is concerned, a “slight chance” of winning \$1 million, is only a gain of \$48! Thus, Alice prefers to buy one ticket rather than none—and rather than two.

Note that this analysis assumes that Alice conceptualizes purchases of more than one lottery ticket as a single transaction; that is, she considers buying two tickets to be an act of paying \$2 for a 0.0001% chance to win, which, as noted, is still just a “slight chance” as far as her language of preference is concerned. One argument for making this assumption is that it allows us to explain with a simple mechanism widely attested behaviour. But in identifying it as an assumption, we can also ask when it does not hold.

The answer to this question has economic implications. If Alice buys a lottery ticket every weekend, we might reasonably assume that she does not lump these purchases together into one cumulative chance of winning. Rather, she considers her purchase of a lottery ticket one week to be distinct from her purchase the previous week. If this conceptual separation of the two purchases were induced by other means, without the time delay, it could be exploited to get people to buy more than one lottery ticket at once. Indeed, this tactic is arguably already in widespread use; for example, the Canadian lottery “Lotto 6/49” offers an option called “Encore” which, for a small additional fee, allow the purchaser to essentially play a second, smaller lottery. ▮

Example 3.4: *Indignant altruism.* Alice and Bob sit down to play a classic game of prisoner’s dilemma, with one twist: neither wishes to live up to low expectations. Specifically, if Bob expects the worst of Alice (i.e. expects her to defect), then Alice, indignant at Bob’s opinion of her, prefers to cooperate. Likewise for Bob. On the other hand, in the absense of such low expectations from their opponent, each will revert to their classical preferences.

The standard prisoner’s dilemma is summarized in Table 6:

	c	d
c	(3,3)	(0,5)
d	(5,0)	(1,1)

Table 6: The classical prisoner’s dilemma.

Let u_A, u_B denote the two players’ utility functions according to this table, and let Γ denote the game form obtained by throwing away these functions: $\Gamma = (\{A, B\}, \Sigma_A, \Sigma_B)$, where $\Sigma_A = \Sigma_B = \{\mathsf{c}, \mathsf{d}\}$. We wish to define an $\mathcal{L}_B(\Phi_\Gamma)$-game that captures the given scenario; to do so we must define new utility functions on $\mathcal{S}$. Informally, if Bob is sure that Alice will defect, then Alice’s utility for defecting is -1, regardless of what Bob does, and likewise reversing the roles of Alice and Bob; otherwise, utility is determined exactly as it is classically.

Formally, we simply define $u'_A : \mathcal{S} \to \mathbb{R}$ by

$$u'_A(S) = \begin{cases} -1 & \text{if } \mathit{play}_A(\mathsf{d}) \in S \text{ and} \\ & B_B\, \mathit{play}_A(\mathsf{d}) \in S \\ u_A(\rho_A(S), \rho_B(S)) & \text{otherwise,} \end{cases}$$

and similarly for u'_B.

Intuitively, cooperating is rational for Alice if she thinks that Bob is sure she will defect, since cooperating in this case would yield a minimum utility of 0, whereas defecting would result in a utility of -1. On the other hand, if Alice thinks that Bob is *not* sure that she will defect, then since her utility in this case is determined classically, it is rational for her to defect, as usual.

This game has much in common with the surprise proposal of Example 1.1: in both games, the essential element is the desire to surprise another player. Perhaps unsurprisingly, when players wish to surprise their opponents, *Nash equilibria* fail to exist—even mixed strategy equilibria. Although we have not yet defined Nash equilibrium in our setting, the classical intuition is wholly applicable: a Nash equilibrium is a state of play where players are happy with their choice of strategies *given accurate beliefs about what their opponents will choose.* But there is a fundamental tension between a state of play where everyone has accurate beliefs, and one where some player successfully surprises another.

We show formally in Section 4.2 that this game has no Nash equilibrium (Proposition 4.2). On the other hand, players can certainly best-respond to their beliefs. In Section 4.3 we provide a natural definition of *rationalizability* in our framework, and show that every strategy for the indignant altruist is rationalizable (Proposition 4.4). ∎

Example 3.5: *The trust game.* Alice is handed \$2 and given a choice: either split the money with Bob, or hand him all of it. If she splits the money, the game is over and they each walk away with \$1. If she hands the money to Bob, it is doubled to \$4, and Bob is offered a choice: either share the money equally with Alice, or keep it all for himself. However, if Bob chooses to keep the money for himself, then he suffers from guilt to the extent that he let Alice down.

This is a paraphrasing of the "psychological trust game" [4]; we consider it here as a normal-form game. The monetary payoffs are summarized in Figure 1:

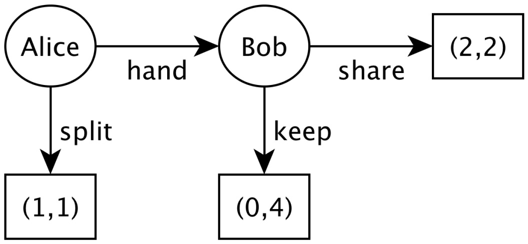

Figure 1: Monetary payoffs in the trust game.

Let m_A and m_B denote the monetary utility functions corresponding to Figure 1, and let $\Gamma = (\{A, B\}, \{\mathsf{split}, \mathsf{hand}\}, \{\mathsf{keep}, \mathsf{share}\})$. To capture Bob's guilt aversion using $\mathcal{L}_B(\Phi_\Gamma)$-situations, let

$$u_B(S) = \begin{cases} -1 & \text{if } \mathit{play}(\mathsf{hand},\ \mathsf{keep}) \in S \\ & \text{and } B_A\, \mathit{play}_B(\mathsf{share}) \in S \\ m_B(\rho_A(S), \rho_B(S)) & \text{otherwise;} \end{cases}$$

Alice's preferences are simply given by

$$u_A(S) = m_A(\rho_A(S), \rho_B(S)).$$

In other words, Bob feels guilty in those situations where Alice hands him the money and is sure he will share it, but he doesn't.[5] On the other hand, even if Alice chooses to hand the money over, u_B tells us that Bob does not feel guilty betraying her provided she had some bit of doubt about his action. We show in Section 4.2 that the only Nash equilibrium in which Alice places any weight at all on her strategy hand is the *pure* equilibrium where she plays hand and Bob plays share (Proposition 4.3).

[5] A subtle issue arises here regarding the sense in which utility is actually "felt". Of course, in a situation where Alice expects Bob to share and he doesn't, he might not, in fact, feel guilty, because he might not realize that Alice expected him to share. In general, if Bob's utility in a given situation is conceptualized as how happy he *actually* feels in that situation, then defining it in terms of something he doesn't have epistemic access to (like Alice's beliefs) is problematic. In fact, this issue arises even in the classical setting: a player may never actually observe the strategy his opponent plays.

One natural reformulation runs as follows: the utility of an outcome for Bob is how happy Bob would feel if he *knew* that was indeed the outcome. While this story seems to do the job in the classical case, it encounters difficulty in the more general context of language-based games because there are situations that Bob can *never* know he is in. For example, any situation S such that $p \wedge \neg B_B p \in S$ is, by construction, not a situation Bob can know he is in. Nonetheless, it makes perfect sense for Bob to prefer such a situation S to some other one S' with, say, $\neg p \wedge B_B \neg p \in S$ (perhaps he prefers to live in world where unicorns in fact exist though he doesn't believe it, rather than discover definitive evidence that they are make-believe).

As is standard, we view utility as a numeric representation of preference, and thinking in terms of preference rather than happiness helps to clarify this issue. It seems to us perfectly reasonable that an agent can contemplate different conceivable situations, which are just descriptions of the world in the language he considers relevant, and assign to them utility values that reflect his preferences among them.

A more satisfying account of this game might involve a finer-grained representation of Alice's expectations. To model this, we must enrich the underlying language. Given a subset $\Theta \subseteq [0,1]$, let $\mathcal{L}_B^\Theta(\Phi_\Gamma)$ be the language obtained by starting with the formulas $play_i(\sigma_i) \in \Phi_\Gamma$ and closing off under $\wedge$, $\neg$, and B_i^θ, where $\theta \in \Theta$. We think of the elements of Θ as indicating "thresholds" or "levels" of belief; the higher the number, the stronger the belief. Semantics for this language are given by augmenting the valuation function as follows:

$$[\![B_i^\theta \varphi]\!] \coloneqq \{\omega \in \Omega \;:\; \mathcal{PR}_i(\omega)([\![\varphi]\!]) \geq \theta\}.$$

Thus, the formula $B_i^\theta \varphi$ is interpreted as saying "player i considers the likelihood of φ to be at least θ". The language $\mathcal{L}_B(\Phi_\Gamma)$ can be viewed as the special case where $\Theta = \{1\}$.

Consider, for example, the language corresponding to the set of thresholds $\Theta = \{1/5, 2/5, 3/5, 4/5, 1\}$. A graded version of Bob's guilt aversion can then be captured in an $\mathcal{L}_B^\Theta(\Phi_\Gamma)$-game by defining $u'_B : \mathcal{S}(\mathcal{L}_B^\Theta(\Phi_\Gamma)) \to \mathbb{R}$ by

$$u'_B(S) = \begin{cases} 4 - k' & \text{if } play(\mathsf{hand},\ \mathsf{keep}) \in S \\ & \text{and } B_A^{1/5}\, play_B(\mathsf{share}) \in S \\ m_B(\rho_A(S), \rho_B(S)) & \text{otherwise,} \end{cases}$$

where

$$k' \coloneqq \max\{k \;:\; B_A^{k/5}\, play_B(\mathsf{share}) \in S\}.$$

As before, Bob feels guilty if he keeps the money that Alice handed to him provided she expected him to share it, but in this case "expected" means "thought there was at least a 20% chance of", and moreover, how guilty Bob feels increases in several discrete increments as Alice's expectations grow stronger.

When $\Theta = [0,1]$, we can define a utility function to capture what might be thought of as "continuous" guilt; that is, guilt that depends in a continuous way on Alice's beliefs: define $u''_B : \mathcal{S}(\mathcal{L}_B^{[0,1]}(\Phi_\Gamma)) \to \mathbb{R}$ by

$$u''_B(S) = \begin{cases} 4 - 5\theta' & \text{if } play(\mathsf{hand},\ \mathsf{keep}) \in S \\ m_B(\rho_A(S), \rho_B(S)) & \text{otherwise,} \end{cases}$$

where

$$\theta' \coloneqq \sup\{\theta \;:\; B_A^\theta\, play_B(\mathsf{share}) \in S\}.$$

In psychological game theory, utility functions depend on beliefs as represented in this continuous manner. We have seen, however, that there are conceptual and theoretical advantages to modeling *categorical* beliefs, where Θ is finite. While it is certainly possible to define a utility function in a psychological game that mimics such categoricity (a step function, for example), such a function is not continuous and therefore not subsumed by the equilibrium analyses that are provided in that literature. Language-based games offer new tools for equilibrium analyses that are able to handle such discontinuities naturally. ∎

Example 3.6: *Pay raise.* Bob has been voted employee of the month at his summer job, an honour that comes with a slight increase (up to \$1) in his per-hour salary, at the discretion of his boss, Alice. Bob's happiness is determined in part by the raw value of the bump he receives in his wages, and in part by the sense of gain or loss he feels by comparing the increase Alice grants him with the minimum increase he expected to get. Alice, for her part, wants Bob to be happy, but this desire is balanced by a desire to save company money.

As usual, we first fix a game form that captures the players and their available strategies. Let $\Gamma = (\{A, B\}, \Sigma_A, \{\cdot\})$, where $\Sigma_A = \{s_0, s_1, \ldots, s_{100}\}$ and s_k represents an increase of k cents to Bob's per-hour salary (Bob has no choice to make, so his strategy set is a singleton). Notice that in this game Bob's preferences depend on his *own* beliefs rather than the beliefs of his opponent. Broadly speaking, this is an example of *reference-dependent preferences*: Bob's utility is determined in part by comparing the actual outcome of the game to some "reference level"—in this case, the minimum expected raise. This game also has much in common with a scenario described by Battigalli and Duwfenberg [4], in which a player Abi wishes to tip her taxi driver exactly as much as he expects to be tipped, but no more.

Define $u_B : \mathcal{S} \to \mathbb{R}$ by setting

$$u_B(S) = k_S + (k_S - r_S),$$

where k_S is the unique integer such that $\mathit{play}_A(s_k) \in S$, and

$$r_S := \min\{r' \; : \; \widehat{B}_B \, \mathit{play}_A(s_{r'}) \in S\}.$$

Observe that r_S is completely determined by Bob's beliefs: it is the lowest raise he considers it possible that Alice will grant him. We think of the first summand k_S as representing Bob's happiness on account of receiving a raise of k_S cents per hour, while the second summand $k_S - r_S$ represents his sense of gain or loss depending on how reality compares to his lowest expectations.

Note that the value of r_S (and k_S) is encoded in S via a finite formula; in other words, we could have written the definition of u_B in a fully expanded form where each utility value is specified by the presence of a formula in S. For instance, the combination $k_S = 5$, $r_S = 2$ corresponds to the formula

$$\mathit{play}_A(s_5) \wedge \widehat{B}_B \, \mathit{play}_A(s_2) \wedge \neg(\widehat{B}_B \, \mathit{play}_A(s_0) \vee \widehat{B}_B \, \mathit{play}_A(s_1))$$

being in S; this combination leads to S having a utility of 8.

Of course, it is just as easy to replace the minimum with the maximum in the above definition (perhaps Bob feels entitled to the most he considers it possible he might get), or even to define the reference level as some more complicated function of Bob's beliefs. The quantity $k_S - r_S$ representing Bob's sense of gain or loss is also easy to manipulate. For instance, given $\alpha, \beta \in \mathbb{R}$, we might define a function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) = \begin{cases} \alpha x & \text{if } x \geq 0 \\ \beta x & \text{if } x < 0, \end{cases}$$

and set

$$u'_B(S) = k_S + f(k_S - r_S).$$

Choosing, say, $\alpha = 1$ and $\beta > 1$ results in Bob's utility u'_B incorporating *loss aversion*: Bob is more upset by a relative loss than he is elated by a same-sized relative gain. These kinds of issues are discussed by Köszegi and Rabin [13]; in Section 5 we analyze a central example from this paper in detail.

Turning now to Alice's preferences, we are faced with a host of modeling choices. Perhaps Alice wishes to grant Bob the smallest salary increase he expects but nothing more. We can capture this by defining $u_A : \mathcal{S} \to \mathbb{R}$ by setting

$$u_A(S) = -|k_S - r_S|.$$

Or perhaps we wish to represent Alice as feeling some fixed sense of guilt if she undershoots, while her disutility for overshooting depends on whether she merely exceeded Bob's lowest expectations, or in fact exceeded even his highest expectations:

$$u'_A(S) = \begin{cases} -25 & \text{if } k_S < r_S \\ r_S - k_S & \text{if } r_S \leq k_S < R \\ r_S - R_S + 2(R_S - k_S) & \text{if } k_S \geq R_S, \end{cases}$$

where

$$R_S := \max\{R' \; : \; \widehat{B}_B \, \mathit{play}_A(s_{R'}) \in S\}.$$

Or perhaps Alice's model of Bob's happiness is sophisticated enough to include his sensations of gain and loss, so that, for example,

$$u''_A(S) = u_B(S) - \delta k_S,$$

where δ is some scaling factor. The framework is rich enough to represent many possibilities. ∎

Example 3.7: *A deeply surprising proposal.* Bob hopes to propose to Alice, but she wants it to be a surprise. He knows that she would be upset if it were not a surprise, so he would prefer not to propose if Alice so much as suspects it. Worse (for Bob), even if Alice does not suspect a proposal, if she suspects that Bob thinks she does, then she will also be upset, since in this case a proposal would indicate Bob's willingness to disappoint her. Of course, like the giant tortoise on whose back the world rests, this reasoning continues "all the way down"...

This example is adapted from a similar example given by Geanakoplos et al. [9]; in their story, the man is considering giving a gift of flowers, but rather than hoping to surprise the recipient, his goal is the exact opposite: to get her flowers just in case she *is* expecting them. Of course, the notion of "expectation" employed, both in their example and ours, is quite a bit more complicated than the usual sense of the word, involving arbitrarily deeply nested beliefs.

Nonetheless, it is relatively painless to represent Bob's preferences in the language $\mathcal{L}_B(\Phi_\Gamma)$, where $\Gamma = (\{A, B\}, \{\cdot\}, \{p, q\})$ and p and q stand for Bob's strategies of proposing and not proposing, respectively (Alice has no decision to make, so her strategy set is a singleton). We use $\widehat{B}_A p$ as our gloss for Alice "so much as suspecting" a proposal. Define $u_B : \mathcal{S} \to \mathbb{R}$ by

$$u_B(S) = \begin{cases} 1 & \text{if } \mathit{play}_B(p) \in S \text{ and} \\ & (\forall k \in \mathbb{N})[\widehat{B}_A(\widehat{B}_B\widehat{B}_A)^k \mathit{play}_B(p) \notin S] \\ 1 & \text{if } \mathit{play}_B(q) \in S \text{ and} \\ & (\exists k \in \mathbb{N})[\widehat{B}_A(\widehat{B}_B\widehat{B}_A)^k \mathit{play}_B(p) \in S] \\ 0 & \text{otherwise,} \end{cases}$$

where $(\widehat{B}_B\widehat{B}_A)^k$ is an abbreviation for $\widehat{B}_B\widehat{B}_A \cdots \widehat{B}_B\widehat{B}_A$ (k times). In other words, proposing yields a higher utility for Bob in the situation S if and only if none of the formulas in the infinite family $\{\widehat{B}_A(\widehat{B}_B\widehat{B}_A)^k \mathit{play}_B(p) \;:\; k \in \mathbb{N}\}$ occur in S.

As in Examples 1.1 and 3.4, and in general when a player desires to surprise an opponent, it is not difficult to convince oneself informally that this game admits no Nash equilibrium. But in this case the infinitary nature of Bob's desire to "surprise" Alice has an even stronger effect: as we show in Section 4.3, no strategy for Bob is even *rationalizable* (Proposition 4.6). ∎

Example 3.8: *Returning a library book.* Alice has learned that a book she borrowed from the library is due back tomorrow. As long as she returns it by tomorrow, she'll avoid a late fee; returning it today, however, is mildly inconvenient.

Here we make use of an extremely simple example to illustrate how to model an ostensibly dynamic scenario in a static framework by employing a suitable underlying language. The idea is straightforward: Alice has a choice to make *today*, but how she feels about it depends on what she might do tomorrow. Specifically, if she returns the library book tomorrow, then she has no reason to feel bad about not returning it today. Since the future has yet to be determined, we model Alice's preferences as depending on what action she takes in the present together with what she *expects* to do in the future.

Let $\Gamma = (A, \{\mathsf{return}, \mathsf{wait}\})$ be a game form representing Alice's two current options, and set $\Phi'_\Gamma := \Phi_\Gamma \cup \{\mathsf{tomorrow}\}$; thus Φ'_Γ is the usual set of primitive propositions (representing strategies) together with a single new addition, $\mathsf{tomorrow}$, read "Alice will return the book tomorrow".

An $\mathcal{L}_B(\Phi'_\Gamma)$-game allows us to specify Alice's utility in a manner consistent with the intuition given above. In particular, we can define $u_A : \mathcal{S}(\mathcal{L}_B(\Phi'_\Gamma)) \to \mathbb{R}$ by

$$u_A(S) = \begin{cases} -1 & \text{if } \mathit{play}_A(\mathsf{return}) \in S \\ 1 & \text{if } \mathit{play}_A(\mathsf{wait}) \wedge B_A\mathsf{tomorrow} \in S \\ -5 & \text{otherwise,} \end{cases}$$

so Alice prefers to wait if she expects to return the book tomorrow, and to return the book today otherwise.

In this example, Alice's utility depends on her beliefs, as it does in psychological game theory. Unlike psychological game theory, however, her utility depends on her beliefs about features of the world aside from which strategies are being played. This is a natural extension of the psychological framework in a language-based setting.

We might want to expand the set of actions by providing Alice with ways to influence her beliefs about tomorrow. For example, perhaps a third strategy is available to her, remind, describing a state of affairs where she keeps the book but places it on top of her keys, thus decreasing the likelihood that she will forget to take it when she leaves the next day. More generally, this simple framework allows us to model *commitment devices* [7]: we can represent players who rationally choose to perform certain actions (like buying a year-long gym membership, or throwing away their "fat jeans") not because these actions benefit them immediately, but because they make it subjectively more likely that the player will perform certain other desirable actions in the future (like going to the gym regularly, or sticking with a diet) that might otherwise be neglected. In a similar manner, we can succinctly capture *procrastination*: if, for example, you believe that you will quit smoking tomorrow, then the health benefits of quitting today instead might seem negligible—so negligible, in fact, that quitting immediately may seem pointless, even foolish. Of course, believing you will do something tomorrow is not the same thing as actually doing it when tomorrow comes, thus certain tasks may be delayed repeatedly. ■

4 Solution Concepts

A number of important concepts from classical game theory, such as *Nash equilibrium* and *rationalizability*, have been characterized epistemically, using Γ-structures. In $\mathcal{L}_B(\Phi_\Gamma)$-games (or, more generally, in language-based games where the language includes belief), we can use these epistemic characterizations to *define* the corresponding solution concepts. This yields natural definitions that generalize those of classical game theory.

4.1 Rationality

A player i is called *rational* if he is best-responding to his beliefs: the strategy σ_i he is using must yield an expected utility that is at least as high as any other strategy σ_i' he could play. In classical game theory, the meaning of this statement is quite clear: player i has beliefs about the strategies his opponents are using in the form of a probability distribution π on Σ_{-i}, and the **expected**

utility of σ_i' is defined to be

$$\sum_{\sigma_{-i} \in \Sigma_{-i}} u_i(\sigma_i', \sigma_{-i}) \cdot \pi(\sigma_{-i}).$$

This definition encodes an important assumption. In order to determine the strategy that maximizes expected utility, players must consider what their expected utility would be if they were to play a different strategy. This, in turn, requires them to have beliefs about what other players would do if they were to play a different strategy. The standard assumption is that a player's beliefs about what other players are doing do not change, regardless of which strategy he is considering. This assumption is easy to overlook, and has received relatively little attention in the literature (but see [11; 21]); however, it is far from innocuous. It rules out, for example, the possibility that players can read each others' body language and thereby glean some information about their opponents' intended strategies.

These issues become much more significant in the context of language-based games. Even if we assume that a player's beliefs about other players' strategies do not change when she contemplates switching to a different strategy, what about her other beliefs? For instance, in Example 3.8, if S is a situation where Alice plays return, what would happen to her beliefs regarding tomorrow if she were to play wait? Should they stay the same? That is far from clear. It seems reasonable to expect that Alice's choice of action should affect her beliefs about when she will return the library book. But answering this question is critical in order to decide if playing wait has higher expected utility than playing return.

In general, it seems that determining what a player's expected utility would be if she were to switch strategies requires more information regarding counterfactuals than is given by a Γ-structure. However, when we restrict our attention to the language $\mathcal{L}_B(\Phi_\Gamma)$, we can make precise the intuition that a player's beliefs about other players' beliefs and strategies remains constant when she contemplates switching strategies. This gives us a general procedure for defining rationality in $\mathcal{L}_B(\Phi_\Gamma)$-games.

A formula $\varphi \in \mathcal{L}_B(\Phi_\Gamma)$ is ***i*-independent** if every occurrence of a subformula of the form $play_i(\sigma_i)$ in φ falls within the scope of some B_j, $j \neq i$. Intuitively, an i-independent formula describes a proposition that is independent of player i's choice of strategy, such as another player's strategy, another player's beliefs, or even player i's beliefs about the other players. Given $S \in \mathcal{S}$, set

$$\rho_{-i}(S) = \{\varphi \in S \,:\, \varphi \text{ is } i\text{-independent}\}.^6$$

Let $\mathcal{S}_{-i}$ denote the image of $\mathcal{S}$ under ρ_{-i}. Elements of $\mathcal{S}_{-i}$ are called ***i*-situations**; intuitively, they are complete descriptions of states of affairs that

[6] As (quite correctly) pointed out by an anonymous reviewer, this notation is not standard, since ρ_{-i} is not a profile of functions of the type ρ_i. Nonetheless, we feel it is appropriate in the sense that, while ρ_i extracts from a given situation player i's strategy, ρ_{-i} extracts "all the rest" (cf. Proposition 4.1), the crucial difference here being that this includes far more than just the strategies of the other players.

are out of player i's control. Informally, an i-situation $S_{-i} \in \mathcal{S}_{-i}$ determines everything about the world (expressible in the language) except what strategy player i is employing. This is made precise in Proposition 4.1. Recall that $\rho_i(S)$ denotes the (unique) strategy that i plays in S, so $play_i(\rho_i(S)) \in S$.

Proposition 4.1: *For each $i \in N$, the map $\vec{\rho}_i : \mathcal{S} \to \Sigma_i \times \mathcal{S}_{-i}$ defined by $\vec{\rho}_i(S) = (\rho_i(S), \rho_{-i}(S))$ is a bijection.*

This identification of $\mathcal{S}$ with the set of pairs $\Sigma_i \times \mathcal{S}_{-i}$ provides a well-defined notion of what it means to alter player i's strategy in a situation S "without changing anything else". By an abuse of notation, we write $u_i(\sigma_i, S_{-i})$ to denote $u_i(S)$ where S is the unique situation satisfying $\vec{\rho}_i(S) = (\sigma_i, S_{-i})$. Observe that for each state $\omega \in \Omega$ and each $i \in N$, there is a unique set $S_{-i} \in \mathcal{S}_{-i}$ such that $\omega \models S_{-i}$. We denote this set by $S_{-i}(M, \omega)$, or just $S_{-i}(\omega)$ when the Γ-structure is clear from context. Then the utility functions u_i induce functions $\hat{u}_i : \Sigma_i \times \Omega \to \mathbb{R}$ defined by

$$\hat{u}_i(\sigma_i, \omega) = u_i(\sigma_i, S_{-i}(\omega)).$$

As in the classical case, we can view the quantity $\hat{u}_i(\sigma_i, \omega)$ as the utility that player i would have if he were to play σ_i at state ω. It is easy to see that this generalizes the classical approach in the sense that it agrees with the classical definition when the utility functions u_i depend only on the outcome.

For each $i \in N$ and $\sigma_i \in \Sigma_i$, provided that $\hat{u}_i(\sigma_i, \cdot\,)$ is measurable, we can define the expected utility of playing σ_i at ω by

$$EU_i(\sigma_i, \omega) := \int_\Omega \hat{u}_i(\sigma_i, \omega')\, d\mathcal{PR}_i(\omega);$$

when Ω is finite, this reduces to

$$EU_i(\sigma_i, \omega) := \sum_{\omega' \in \Omega} \hat{u}_i(\sigma_i, \omega') \cdot \mathcal{PR}_i(\omega)(\omega').$$

We can then define $BR_i : \Omega \to 2^{\Sigma_i}$ by

$$BR_i(\omega) = \{\sigma_i \in \Sigma_i \,:\, (\forall \sigma_i' \in \Sigma_i)[EU_i(\sigma_i, \omega) \geq EU_i(\sigma_i', \omega)]\};$$

thus $BR_i(\omega)$ is the set of *best-reponses* of player i to his beliefs at ω, that is, the set of strategies that maximize his expected utility.

With this apparatus in place, we expand the underlying language to incorporate *rationality* as a formal primitive. Note that we are *not* replacing $\mathcal{L}_B(\Phi_\Gamma)$ as the underlying language of the game over which the utility functions are defined, but simply defining a richer language that will be useful for analyzing the game. Let

$$\Phi_\Gamma^{rat} := \Phi_\Gamma \cup \{RAT_i \,:\, i \in N\},$$

where we read RAT_i as "player i is rational". We also employ the syntactic abbreviation $RAT \equiv RAT_1 \wedge \cdots \wedge RAT_n$. Intuitively, $\mathcal{L}_B(\Phi_\Gamma^{rat})$ allows us to reason about whether or not players are being rational with respect to their beliefs and preferences, in the sense of expected utility maximization. Formally, we extend the valuation function $\llbracket \cdot \rrbracket_M$ to $\mathcal{L}_B(\Phi_\Gamma^{rat})$ by setting

$$\llbracket RAT_i \rrbracket_M \quad := \quad \{\omega \in \Omega \, : \, s_i(\omega) \in BR_i(\omega)\}.$$

Thus RAT_i holds at state ω just in case the strategy that player i is playing at that state, $s_i(\omega)$, is a best-response to his beliefs.[7]

4.2 Nash equilibrium

Having formalized rationality, we are in a position to draw on work that characterizes solutions concepts in terms of RAT.

Let $\Gamma = (N, (\Sigma_i)_{i \in N})$ be a game form in which each set Σ_i is finite, and let $\Delta(\Sigma_i)$ denote the set of all probability measures on Σ_i. Elements of $\Delta(\Sigma_i)$ are the **mixed strategies** of player i. Given a *mixed strategy profile*

$$\mu = (\mu_1, \ldots, \mu_n) \in \Delta(\Sigma_1) \times \cdots \times \Delta(\Sigma_n),$$

we define a Γ-structure M_μ that, in a sense made precise below, captures "equilibrium play" of μ and can be used to determine whether or not μ constitutes a Nash equilibrium.

Set

$$\Omega_\mu = supp(\mu_1) \times \cdots \times supp(\mu_n) \subseteq \Sigma_1 \times \cdots \times \Sigma_n.$$

For each $\sigma, \sigma' \in \Omega_\mu$, let

$$\mathcal{PR}_{\mu,i}(\sigma)(\sigma') = \begin{cases} \prod_{j \neq i} \mu_j(\sigma_j) & \text{if } \sigma_i = \sigma'_i \\ 0 & \text{otherwise.} \end{cases}$$

Let $M_\mu = (\Omega_\mu, id_{\Omega_\mu}, \vec{\mathcal{PR}}_\mu)$. It is easy to check that M_μ is a Γ-structure; call it the **characteristic Γ-structure for μ**. At each state in M_μ, each player i is sure of his own strategy and has uncertainty about the strategies of his opponents; however, this uncertainty takes the form of a probability distribution

[7] There is a subtlety here. Normally, we define the valuation function $\llbracket \varphi \rrbracket_M$ (or, equivalently, $\models$) by induction on the structure of φ. But here it is important that we define $\llbracket RAT_i \rrbracket_M$ *after* we have defined $\llbracket \varphi \rrbracket_M$ for all formulas in $\mathcal{L}_B(\Phi_\Gamma)$. The semantics of RAT_i implicitly assumes this, since it depends on the function $\hat{u}_i$, which in turn depends on the $\mathcal{L}_B(\Phi_\Gamma)$-formulas that are satisfied at each state. Moreover, had we added the formulas RAT_i to the underlying language there would have been circularity in the semantics: to define rationality, we would need to define best response, while to define best response, we would need to define the utility function on situations that included formulas that talk about rationality. Nevertheless, it does not seem so unreasonable to have preferences that depend on rationality. For example, a player might prefer to have others believe that he is irrational, and therefore might play an arguably incredible threat. We defer a discussion of these issues to future work.

weighted according to μ_{-i}, so in effect each player i correctly ascribes the mixed strategy μ_j to each of his opponents $j \neq i$. It is well known (and easy to show) that a mixed strategy profile μ is a Nash equilibrium in the classical sense if and only if each player is rational (i.e. maximizing expected utility) at every state in the characteristic Γ-structure for μ. Accordingly, we *define* a **Nash equilibrium** (in an $\mathcal{L}_B(\Phi_\Gamma)$-game) to be a mixed strategy profile μ such that $M_\mu \models RAT$. It is immediate that this definition generalizes the classical definition of Nash equilibrium.

We note that there are other epistemic characterizations of Nash equilibrium besides the one presented here (see, e.g., [3], which focuses on the role of a common prior and common knowledge of "conjectures"). While in the classical setting they all generate equivalent solution concepts, this may not be the case in our more general model. We believe that investigating the solution concepts that arise by teasing apart such classically equivalent notions is an interesting and promising direction for future research.

In contrast to the classical setting, Nash equilibria are not guaranteed to exist in general; indeed, this is the case for the indignant altruism game of Example 3.4.

Proposition 4.2: *There is no Nash equilibrium in the indignant altruism game.*

Proof: We must show that for every mixed strategy profile

$$\mu = (\mu_A, \mu_B) \in \Delta(\{\mathsf{c}, \mathsf{d}\}) \times \Delta(\{\mathsf{c}, \mathsf{d}\}),$$

the corresponding characteristic Γ-structure $M_\mu \not\models RAT$.

Suppose first that $\mu_A(\mathsf{c}) > 0$. Then $M_\mu \models \neg B_B \, play_A(\mathsf{d})$, which implies that Alice's utility at every state in M_μ coincides with the classical prisoner's dilemma, so she is not rational at any state where she cooperates. Since, by definition, M_μ contains a state where Alice cooperates, we conclude that $M_\mu \not\models RAT_A$, so μ cannot be a Nash equilibrium.

Suppose instead that $\mu_A(\mathsf{c}) = 0$. Then $M_\mu \models B_B \, play_A(\mathsf{d})$, and so Alice, being sure of this, is not rational at any state where she defects, since by definition she is guaranteed a utility of -1 in that case. By definition, M_μ contains a state where Alice defects (in fact, Alice defects in every state), so we can conclude as above that $M_\mu \not\models RAT_A$, which means that μ cannot be a Nash equilibrium. ∎

Why does Nash equilibrium not exist in this example? Roughly speaking, the utility functions in this game exhibit a kind of "discontinuity": the utility of defecting is -1 precisely when your opponent is 100% certain that you will defect. However, as soon as this probability dips below 100%, *no matter how small the drop*, the utility of defecting jumps up to at least 1.

Broadly speaking, this issue arises in $\mathcal{L}$-games whenever $\mathcal{L}$ expresses a coarse-grained notion of belief, such as the underlying language in this example, which only contains belief modalities representing 100% certainty. However, since coarseness is a central feature we wish to model, the lack of existence of Nash equilibria in general might be viewed as a problem with the notion of *Nash equilibrium* itself, rather than a defect of the underlying language. Indeed, the requirements that a mixed strategy profile must satisfy in order to qualify as a Nash equilibrium are quite stringent: essentially, each player must evaluate his choice of strategy *subject to the condition that his choice is common knowledge*! As we have seen, this condition is not compatible with rationality when a player's preference is to do something unexpected.

More generally, this tension arises with any solution concept that requires players to have common knowledge of the mixed strategies being played (the "conjectures", in the terminology of Aumann and Brandenburger [3]). In fact, Proposition 4.2 relies only on second-order knowledge of the strategies: whenever Alice knows that Bob knows her play, she is unhappy. In particular, any alternative epistemic characterization of Nash equilibrium that requires such knowledge is subject to the same non-existence result. Furthermore, we can use the same ideas to show that there is no *correlated equilibrium* [2] in the indignant altruism game either (once we extend correlated equilibrium to our setting); this follows from the fact that in a correlated equilibrium players must still have *correct* beliefs about the strategies their opponents might play, and these beliefs are common knowledge.

All this is not to say that Nash equilibrium is a useless concept in this setting, but merely that we should not expect a general existence theorem in the context of belief-dependent preferences with coarse beliefs. For an example of an $\mathcal{L}_B(\Phi_\Gamma)$-game in which Nash equilibria exist and are informative, we examine again the "trust game" of Example 3.5.

Proposition 4.3: *In the trust game, the only Nash equilibrium in which Alice places positive weight on* hand *is the pure equilibrium* $(\mathsf{hand}, \mathsf{share})$.

Proof: Suppose that

$$\mu = (\mu_A, \mu_B) \in \Delta(\{\mathsf{split}, \mathsf{hand}\}) \times \Delta(\{\mathsf{keep}, \mathsf{share}\})$$

is a Nash equilibrium with $\mu_A(\mathsf{hand}) > 0$. Then there is some state $\omega \in M_\mu$ at which Alice is rationally playing hand. Since Alice can rationally play hand only if she believes with sufficient probability that Bob is playing share, there must be some state $\omega' \in M_\mu$ at which Bob is playing share. Moreover, since by assumption $M_\mu \models RAT$, we know that at ω' Bob is *rationally* playing share. But Bob can rationally play share only if he believes with sufficient probability that $B_A\, play_B(\mathsf{share})$ holds. However, by definition of M_μ, if $B_A\, play_B(\mathsf{share})$ holds at *any* state, then it must hold at *every* state because in this case $\mu_B(\mathsf{share}) = 1$, on account of the fact that in a Nash equilibrium players' beliefs about the strategies of their opponents are always correct.

It is easy to see that when $\mu_B(\mathsf{share}) = 1$, the only rational play for Alice in M_μ is hand, and that when $\mu_A(\mathsf{hand}) = \mu_B(\mathsf{share}) = 1$, we have $M_\mu \models RAT$. This establishes the desired result. ▮

4.3 Rationalizability

In this section, we define rationalizability in language-based games in the same spirit as we defined Nash equilibrium in Section 4.2: epistemically. As shown by Tan and Werlang [22] and Brandenburger and Dekel [6], common belief of rationality characterizes rationalizable strategies. Thus, we define rationalizability that way here.

Let $\mathcal{L}_{CB}(\Phi_\Gamma^{rat})$ be the language generated by starting with the formulas in Φ_Γ and closing off under $\wedge$, $\neg$, the unary operators B_i, $i = 1, \ldots, n$, and CB. We read $CB\varphi$ as "there is common belief of φ". Extend $[\![\cdot]\!]_M$ to $\mathcal{L}_{CB}(\Phi_\Gamma^{rat})$ by setting

$$[\![CB\varphi]\!]_M \quad := \quad \bigcap_{k=1}^{\infty} [\![EB^k\varphi]\!]_M,$$

where

$$\begin{aligned} EB\varphi \quad &\equiv \quad B_1\varphi \wedge \cdots \wedge B_n\varphi, \text{ and} \\ EB^k\varphi \quad &\equiv \quad EB(EB^{k-1}\varphi). \end{aligned}$$

For convenience, we stipulate that $EB^0\varphi \equiv \varphi$. We read $EB\varphi$ as "everyone believes φ". Thus, intuitively, $CB\varphi$ holds precisely when everyone believes φ, everyone believes that everyone believes φ, and so on. We define a strategy $\sigma_i \in \Sigma_i$ to be **rationalizable** (in an $\mathcal{L}_B(\Phi_\Gamma)$-game) if the formula $play_i(\sigma_i) \wedge CB(RAT)$ is satisfiable in some Γ-structure.

Although there are no Nash equilibria in the indignant altruism game, as we now show, every strategy is rationalizable.

Proposition 4.4: *Every strategy in the indignant altruism game is rationalizable.*

Proof: Consider the Γ-structure in Figure 2.

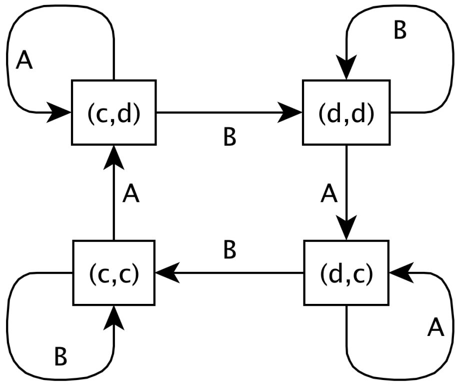

Figure 2: A Γ-structure for indignant altruism.

The valuations of the primitive propositions at each of the four states are labeled in the obvious way. Arrows labeled i based at state ω should be interpreted as pointing to all and only those states ω' such that $\mathcal{PR}_i(\omega)(\omega') > 0$ (in particular, in this example, every probability measure assigns probability 1 to some single state).

As discussed in Example 3.4, it is rational to cooperate in this game if you believe that your opponent believes that you will defect, and it is rational to defect if you believe that your opponent believes you will cooperate. Given this, it is not difficult to check that RAT holds at each state of this Γ-structure; therefore, so does $CB(RAT)$. Thus, by definition, every strategy is rationalizable. ▮

Does every language-based game admit a rationalizable strategy? Every classical game does. This follows from the fact that every strategy in a Nash equilibrium is rationalizable, together with Nash's theorem that every (finite) game has a Nash equilibrium (cf. [19]). In the language-based setting, while it is immediate that every strategy in a Nash equilibrium is rationalizable, since Nash equilibria do not always exist, we cannot appeal to this argument.

In the classical setting, the existence of rationalizable strategies can also be proved by defining a certain iterative deletion procedure and showing that it always terminates in a nonempty set of strategy profiles, and that these profiles are precisely the rationalizable ones. We provide a natural condition that guarantees that this type of approach also works for language-based games. Moreover, we show by example that when this condition does not hold, the existence of rationalizable strategies is not guaranteed.

Perhaps the most straightforward kind of deletion procedure one might propose in our setting works roughly as follows: consider the set of all states in all Γ-structures. Mark those states that fail to satisfy RAT. Next, mark those states ω that include an already-marked state in the support of one of the player's probability measures $\mathcal{PR}_i(\omega)$. These are the states that fail to satisfy $EB(RAT)$. Iterating this process, it is not difficult to see that the states marked at the kth step are those that fail to satisfy $EB^k(RAT)$. Thus, the only states that are never marked are those that satisfy $CB(RAT)$. Moreover, the following lemma (which will play an important role later) implies that at each *finite* stage of this procedure, we are left with a nonempty set of unmarked states.

Lemma 4.5: *$EB^k(RAT)$ is satisfiable for all $k \in \mathbb{N}$.*

Unfortunately, it is not true in general that this procedure always terminates after a finite number of iterations, nor is it clear how to go about showing that any states remain unmarked in the limit, without already knowing that $CB(RAT)$ is satisfiable. The problem here seems to be the unwieldy nature of "the set of all states in all Γ-structures". We therefore work with what is essentially a projection of this set: the set of all situations.

Given any language $\mathcal{L}$, we can topologize $\mathcal{S}(\mathcal{L})$ by taking as basic open sets the collection $\{U_\varphi \,:\, \varphi \in \mathcal{L}\}$, where $U_\varphi := \{S \in \mathcal{S}(\mathcal{L}) \,:\, \varphi \in S\}$. Thus, two situations are in the same open set U_φ just in case they both contain the formula φ; intuitively, two situations are "close" if they have many formulas in common.

Given a set of formulas F and a formula φ, we write $F \models \varphi$, and say that F **entails** φ, if every state that satisfies F also satisfies φ; in other words, F entails φ when $F \cup \{\neg\varphi\}$ is not satisfiable. A logic is said to be **compact** if, whenever $F \models \varphi$, there is some finite subset $F' \subseteq F$ such that $F' \models \varphi$.[8]

It is straightforward to check that $\mathcal{S}(\mathcal{L})$ is compact (as a topological space) just in case $\mathcal{L}$ is compact (as a logic). Furthermore, it is well-known that the KD45 belief logic is compact [5]. Unfortunately, compactness is not necessarily preserved when we augment the logic with primitive propositions RAT_i as in Section 4.1—a player may fail to be rational for an "infinitary" reason. Take, for instance, the deeply surprising proposal of Example 3.7. It is not hard to see that

$$\{play_B(q)\} \cup \{B_B \neg \widehat{B}_A (\widehat{B}_B \widehat{B}_A)^k play_B(p) \,:\, k \in \mathbb{N}\} \models \neg RAT_B.$$

However, no finite subset of this collection is sufficient to entail Bob's irrationality: there will always be some k so high that, should Alice "expect" a proposal at this kth order of "expectation", Bob is indeed rational not to propose. Games with this type of infinitary structure can fail to have rationalizable strategies.

Proposition 4.6: *The deeply surprising proposal game has no rationalizable strategies.*

Proof: Fix a Γ-structure $M = (\Omega, (s_i)_{i \in N}, (\mathcal{PR}_i)_{i \in N})$ and suppose for contradiction that $\omega \in \Omega$ is such that $\omega \models CB(RAT)$. Let "expect*" denote the

[8] Equivalently, for every set of formulas F, F is satisfiable if and only if every finite subset of F is satisfiable.

infinitary notion of expectation at play in this example, and consider first the case where Alice does not *expect** a proposal at state ω: that is, for all $k \in \mathbb{N}$, $\omega \models \neg \widehat{B}_A(\widehat{B}_B\widehat{B}_A)^k \mathit{play}_B(p)$. Then, for all $k \in \mathbb{N}$, $\omega \models B_A(B_B B_A)^k \neg \mathit{play}_B(p)$; taking $k = 0$, it follows that $\mathcal{PR}_A(\omega)([\![\neg \mathit{play}_B(p)]\!]_M) = 1$. Moreover, since $CB(RAT)$ holds at ω, we know in fact that $\mathcal{PR}_A(\omega)([\![\neg \mathit{play}_B(p) \wedge RAT_B]\!]_M) = 1$. But if Bob is rationally *not* proposing at a state ω', then he must at least consider it possible that Alice expects* a proposal. That is, for each $\omega' \in [\![\neg \mathit{play}_B(p) \wedge RAT_B]\!]_M$, there is a k such that $\omega' \models \widehat{B}_B(\widehat{B}_A(\widehat{B}_B\widehat{B}_A)^k \mathit{play}_B(p))$. Thus, $\mathcal{PR}_A(\omega)(\cup_{k=0}^{\infty}[\![(\widehat{B}_B\widehat{B}_A)^{k+1}]\!]_M) = 1$. Since $\mathcal{PR}_A(\omega)$ is countably additive, it follows that there is a k such that $\mathcal{PR}_A(\omega)([\![(\widehat{B}_B\widehat{B}_A)^{k+1}\mathit{play}_B(p)]\!]_M) > 0$. Hence, $\omega \models \widehat{B}_A(\widehat{B}_B\widehat{B}_A)^{k+1}\mathit{play}_B(p)$, contradicting our original assumption. Thus we have shown that any state where $CB(RAT)$ holds is a state where Alice expects* a proposal.

So suppose now that Alice expects* a proposal at ω. It follows that there exists some $\omega' \in \Omega$ with $\omega' \models \mathit{play}_B(p) \wedge CB(RAT)$. But if Bob is rationally playing p at ω', he must consider it possible that Alice doesn't expect* it; from this it follows that there exists a state $\omega'' \in \Omega$ with $\omega'' \models CB(RAT)$ but where Alice doesn't expect* a proposal, which we have seen is impossible.

This completes the argument: $CB(RAT)$ is not satisfiable. It is worth noting that this argument fails if we replace "expects*" with "expects$^{\leq K}$", where this latter term is interpreted to mean $(\forall k \leq K)[\neg \widehat{B}_A(\widehat{B}_B\widehat{B}_A)^k \mathit{play}_B(p)]$. ∎

We now provide a condition that guarantees the existence of rationalizable strategies:

(**CR**) For all $S \in \mathcal{S}$, if $S \models \neg RAT$ then there is a finite subset $F \subset S$ such that $F \models \neg RAT$.

Theorem 4.7: *(CR) implies that rationalizable strategies exist.*

We think of $S \models \neg RAT$ as saying that the situation S is not *compatible with rationality*: there is no state satisfying S at which RAT_i holds for each player i. Property (CR) then guarantees that there is some "finite witness" $F \subset S$ to this fact. In other words, given any situation not compatible with rationality, there is a finite description of that situation that ensures this incompatibility. Note that the deeply surprising proposal game fails to satisfy (CR).

How stringent of a requirement is the condition (CR)? A partial answer to this question is given by the following proposition.

Proposition 4.8: *Every finitely-specified $\mathcal{L}_B(\Phi_\Gamma)$-game satisfies (CR).*

Corollary 4.9: *Every finitely-specified $\mathcal{L}_B(\Phi_\Gamma)$-game has a rationalizable strategy.*

Since we expect to encounter finitely-specified games most often in practice, this suggests that the games we are likely to encounter will indeed have rationalizable strategies.

5 Case Study: Shopping for Shoes

In this section we take an in-depth look at an example that Kőszegi and Rabin [13] (henceforth KR) analyze in detail: shopping for shoes. KR apply their theory of reference-dependent preferences to study a typical consumer's decision-making process, illustrating several insights and predictions of their formalism along the way. We do the same, modeling the interaction as an $\mathcal{L}_B(\Phi_\Gamma)$-game and comparing this approach to that of KR. The development in this section can easily be generalized to more a refined language such as $\mathcal{L}_B^{\Theta}(\Phi_\Gamma)$; however, we choose to work with a minimal language in order to make clear the surprising richness that even the coarsest representation of belief can exhibit.

5.1 Setup

The game form $\Gamma = (\{C, R\}, \Sigma_C, \Sigma_R)$ consists of two players: a consumer C and a retailer R. As we are interested only in the consumer's decisions and motivations, we model the retailer's preferences with a constant utility function; in essence, R plays the role of "the environment".

Let Σ_R be a set of non-negative real numbers, the *prices*; $p \in \Sigma_R$ represents the retailer setting the price of a pair shoes to be p units. The consumer's choice is essentially whether or not to buy the given pair of shoes. However, since we model play as simultaneous, and whether or not C decides to buy might depend on what R sets the price at, the strategies available to C should reflect this. Let Σ_C be a set of real numbers, the *thresholds*; $t \in \Sigma_C$ represents the threshold cost at which C is no longer willing to buy the shoes. An outcome of this game is therefore a threshold-price pair $(t, p) \in \Sigma$; intuitively, the shoes are purchased for price p if and only if $t > p$.

The consumer's utility depends on the outcome of the game together with a "reference level". A reference level is like an imaginary outcome that the actual outcome of the game is compared to, thereby generating sensations of gain or loss. Roughly speaking, KR interpret the reference level as being determined by a player's expectations, that is, her (probabilistic) beliefs about outcomes. Formally, they allow for stochastic reference levels given by probability measures on the set of outcomes; sensations of gain or loss with respect to stochastic reference levels are calculated by integrating with respect to these probability measures. By contrast, in our framework, beliefs can affect utility only insofar as they can be expressed in the underlying language. The coarseness of the language $\mathcal{L}_B(\Phi_\Gamma)$ is therefore a departure from KR's approach; nonetheless, we will see that many of their insights also arise in our framework in a coarse setting (and, of course, we can reproduce their results with a richer language).

To clarify our definition of utility as well as to conform to the exposition given by KR as closely as possible, we begin by defining some auxiliary functions. Following KR, we think of the outcome of the game as far as utility is concerned as being divided into two dimensions, the first tracking the money spent, and the second tracking the product obtained. As a separate matter, we also think of utility itself as coming in two components: *consumption utility*, which is akin to the usual notion in classical game theory depending solely on the outcome, and *gain-loss utility*, the component that depends on the reference level.

The two dimensions of consumption utility are given by functions $m_i : \Sigma \to \mathbb{R}$ defined by

$$m_1(t,p) = \begin{cases} -p & \text{if } p < t \\ 0 & \text{if } p \geq t \end{cases}$$

and

$$m_2(t,p) = \begin{cases} 1 & \text{if } p < t \\ 0 & \text{if } p \geq t. \end{cases}$$

As KR do, we assume *additive separability* of consumption utility, so the function $m = m_1 + m_2$ gives C's total consumption utility. This function captures the intuition that, when the price of the shoes is below the threshold for purchase, C buys the shoes and therefore gets a total consumption utility of $1 - p$: a sum of the "intrinsic" value of the shoes to her (normalized to 1), and the loss of the money she paid for them $(-p)$. Otherwise, C neither spends any money nor gets any shoes, so her utility is 0.

Next we define functions representing the two corresponding dimensions of gain-loss utility, $n_i : \Sigma^2 \to \mathbb{R}$, by

$$n_i(t,p \,|\, s,q) = \mu(m_i(t,p) - m_i(s,q)),$$

where $\mu : \mathbb{R} \to \mathbb{R}$ is a fixed function that we discuss shortly. The value $n_i(t,p \,|\, s,q)$ should be thought of as the gain-loss utility (in dimension i) of the outcome (t,p) given the reference level (s,q). Furthermore, as KR do, we assume that gain-loss utility is a function of the difference between the consumption utility of the actual outcome, $m_i(t,p)$, and the consumption utility of the reference level, $m_i(s,q)$. Following KR, for the purposes of this example we let

$$\mu(x) = \begin{cases} \eta x & \text{if } x > 0 \\ \lambda \eta x & \text{if } x \leq 0, \end{cases}$$

where $\eta < 0$ and $\lambda > 1$. Thus, λ implements loss-aversion by ensuring that any sense of loss is λ-times greater than the positive feeling associated with a corresponding gain.

As with consumption utility, we assume that gain-loss utility is additively separable, so the function $n = n_1 + n_2$ gives the total gain-loss utility. Finally, C's total utility $u : \Sigma^2 \to \mathbb{R}$ is given by

$$u(t, p \,|\, s, q) = m(t, p) + n(t, p \,|\, s, q),$$

the sum of her total consumption utility and her total gain-loss utility.

As mentioned, KR interpret the reference level as being determined by beliefs; indeed, this is the foundation of one of the main contributions of their paper. We might therefore model C's reference level as being entirely determined by her first-order beliefs about outcomes; for the time being, we adopt this modeling assumption, although we explore a different option in Section 5.3. Note that under this assumption, in our framework a reference level (s, q) must satisfy $s = t$, where t is the actual threshold chosen by C; this follows from the fact that players are always sure of their own strategies. Thus, C's reference level is completely captured by the value q, namely, what she thinks the price will be set at.

Having formalized a notion of utility comparing an outcome to a single reference level, we must extend this to account for uncertainty on the part of the consumer. In other words, if a reference level is conceptualized as an expected outcome, we must specify C's utility when she considers more than one outcome possible.

Let $\mathit{ref}_C : \mathcal{S}(\mathcal{L}_B(\Phi_\Gamma)) \to 2^{\Sigma_R}$ be defined by

$$\mathit{ref}_C(S) = \{q \in \Sigma_R \;:\; \widehat{B}_C \, \mathit{play}_R(q) \in S\}.$$

This function extracts from a given $\mathcal{L}_B(\Phi_\Gamma)$-situation S the set of all prices $q \in \Sigma_R$ such that C considers it possible that R might play q. This set plays the same role for us that a stochastic reference level G plays for KR; in a sense, $\mathit{ref}_C(S)$ is the support of a distribution like G.

To incorporate the uncertainty expressed by the stochastic beliefs G into a measure of utility, KR integrate u against G, yielding in essence a weighted average. We can bypass the calculus and just take the average, defining $u_C : \mathcal{S}(\mathcal{L}_B(\Phi_\Gamma)) \to \mathbb{R}$ by

$$u_C(S) = |\mathit{ref}_C(S)|^{-1} \sum_{q \in \mathit{ref}_C(S)} u(t, p \,|\, t, q),$$

where $t = \rho_C(S)$ and $p = \rho_R(S)$ are the strategies actually played by C and R in the situation S, respectively.

Of course, this is far from the only way in which we might massage the set $\mathit{ref}_C(S)$ into a utility function for C; for instance, analogously to the "pay raise" of Example 3.6, we might stipulate that C's reference level is given by her highest price expectation:

$$u'_C(S) = u(t, p \,|\, t, \max(\mathit{ref}_C(S))).$$

In order to parellel the definitions of KR as closely as possible, however, we focus on utility as given by averaging reference levels.

5.2 Predictions

The game form Γ, equipped with the utility function u_C (as well as a constant utility function u_R), forms an $\mathcal{L}_B(\Phi_\Gamma)$-game. We now demonstrate that, despite the coarseness of the underlying language, important predictions from KR's framework persist. Notably, we accomplish this without making use of the solution concepts that they define, but instead with a basic assumption of rationality on the part of the consumer (as in Section 4.1). In Section 5.3, we explore KR's solution concepts of *personal equilibrium* and *preferred personal equilibrium* in some detail.

We begin by considering the consumer's behaviour under price certainty. KR show that, in this case, the consumer's preferred personal equilibrium is to buy the shoes if the cost is below their intrisic value, $p < 1$, and not to buy the shoes when $p > 1$.

Fix a Γ-structure M and suppose that ω is a state at which C is certain that the shoes will be offered for price p:

$$\mathcal{PR}_C(\omega)([\![play_R(p)]\!]_M) = 1.$$

A rational consumer, by definition, seeks to maximize expected utility; in this case, as she has no doubt about the price of the shoes, her expected utility on playing $t \in T$ is simply $u(t, p \,|\, t, p)$. This is because in every state she considers possible both the actual price and the expected price are p. More formally, for every $\omega' \in \mathcal{PR}_C[\omega]$ we know that $ref_C(S(\omega')) = \{p\}$, and therefore

$$\hat{u}_C(t, \omega') = u(t, p \,|\, t, p) = \begin{cases} 1 - p & \text{if } p < t \\ 0 & \text{if } p \geq t. \end{cases}$$

It follows that in the absence of price uncertainty, a rational consumer chooses a threshold $t > p$ (that is, chooses to buy the shoes at the expected price) whenever $p < 1$, and chooses a threshold $t \leq p$ whenever $p > 1$; for instance, choosing $t = 1$ accomodates both of these restrictions at once. Thus, in this model, when a rational consumer is certain of the price, sensations of gain or loss do not enter into the picture.

Next we consider a case of price uncertainty. Fix a Γ-structure M and suppose that ω is a state at which C is considers it possible that the shoes will be offered at one of two prices: p_L and p_M, where $p_L < p_M$. In other words, $ref_C(S(\omega)) = \{p_L, p_M\}$. Suppose also that $T = \{t_L, t_H\}$, where $p_L < t_L < p_M < t_H$. Thus, the two strategies available to C constitute a choice between buying at price p_M or not, while buying at price p_L is a foregone conclusion. As we saw, if the consumer were certain that the price would be p_M, she could rationally play t_H just in case $p_M \leq 1$. Under uncertainty, however, the rational threshold for buying can change.

By definition, C's expected utility is some convex combination of her utility in case R plays p_M and her utility in case R plays p_L. We analyze each case in turn.

First consider the case where R plays p_M. Then C's utility for playing t_L is equal to

$$m(t_L, p_M) + \frac{1}{2}[n(t_L, p_M \mid t_L, p_L) + n(t_L, p_M \mid t_L, p_M)],$$

her consumption utility m plus the average gain-loss utility for the two reference levels she considers possible. This evaluates to

$$0 + \frac{1}{2}[\mu(0 - (-p_L)) + \mu(0 - 1) + 0] = \frac{\eta p_L - \lambda\eta}{2}.$$

Similarly, C's utility for playing t_H is

$$m(t_H, p_M) + \frac{1}{2}[n(t_H, p_M \mid t_H, p_L) + n(t_H, p_M \mid t_H, p_M)],$$

which evaluates to

$$1 - p_M + \frac{-\lambda\eta(p_M - p_L)}{2}.$$

It follows that playing t_H yields a higher payoff than playing t_L precisely when

$$p_M < 1 + p_L \cdot \frac{\eta(\lambda - 1)}{2 + \lambda\eta}.$$

In the case where R plays p_L, analogous calculations show that t_H is preferred to t_L precisely when

$$p_M > 1 - p_L(\lambda - 1).$$

Since, as noted above, C's expected utility at ω is some convex combination of her utility in the two cases just analysed, we can see that whenever

$$1 - p_L(\lambda - 1) < p_M < 1 + p_L \cdot \frac{\eta(\lambda - 1)}{2 + \lambda\eta}, \tag{1}$$

expected utility is maximized by choosing t_H. In particular, buying the shoes for a price $p_M > 1$ can be rational; moreover, the extra amount $p_M - 1$ that it is always rational to pay is determined by the upper bound of the inequality (1), which is increasing in p_L. Intuitively, the higher the price p_L the consumer was willing to buy the shoes at no matter what, the less of a loss it feels like to pay a little bit extra. Equivalently, the lower the price p_L, the more of a loss it feels like by comparison to pay the higher price p_M. This is the "comparison effect" found by KR.

5.3 Intention

As we have seen, under price certainty, the consumer cannot rationally purchase the shoes if they are being offered at a price $p > 1$. This corresponds to a prediction of KR: in their terminology, buying if $p < 1$ and only if $p \leq 1$ is the unique *preferred personal equilibrium* under price certainty. However, the weaker of the two solution concepts they propose tells a different story. Still assuming price certainty, KR show that both buying for sure and not buying for sure (provided the price is not *too* high or low) are personal equilibria for the consumer.

The idea has a certain appeal: if the consumer is somehow set on a purchase, then a failure to follow through might generate a sense of loss that can overcome a certain amount of overcharging. In essence, people will pay extra to avoid disappointment. Similarly, according to KR, people will pass up a good deal if they had their mind set in advance on saving their money.

KR work in a dynamic setting where this intuition can be cashed out temporally. First, the consumer forms an expectation that she will buy the shoes before she even gets to the store. Upon arrival, she realizes (say) that they are more expensive than she had thought, and updates her beliefs accordingly. However, crucially, she *does not update her reference level* vis-a-vis her intention to buy. Intuitively, as far as being disappointed goes, her reference level is determined by her *old* expectation to buy. Indeed, when unexpected calamity or fortune befalls someone, they typically do not update their expectations immediately and proceed as if the status quo has merely been maintained.

In what follows, we sketch a formalism within which we can tell this type of story; in keeping with the theme of this work, the idea boils down to the right choice of underlying language. Notably, however, the language we employ is not fundamentally temporal in nature. This suggests, we feel, that the corresponding notion at play in KR's work, although presented in a dynamic setting, is better viewed as an instance of a more general construction. We call it *intention*.

Let

$$\Phi_\Gamma^{int} = \Phi_\Gamma \cup \{int_i(\sigma_i) \,:\, i \in N, \sigma_i \in \Sigma_i\},$$

and consider the language $\mathcal{L}_B(\Phi_\Gamma^{int})$. We read $int_i(\sigma_i)$ as "player i intends to play σ_i". An *intentional* Γ*-structure* is a Γ-structure M equipped with additional functions $\iota_i : \Omega \to \Sigma_i$ called the *intention functions* such that whenever $\omega' \in \mathcal{PR}_i[\omega]$, we have $\iota_i(\omega') = \iota_i(\omega)$. This condition ensures that each player is sure of his own intentions. A valuation function $[\![\cdot]\!]_M$ is defined recursively on $\mathcal{L}_B(\Phi_\Gamma^{int})$ as before, with the additional clause

$$[\![int_i(\sigma_i)]\!]_M \coloneqq \{\omega \in \Omega \,:\, \iota_i(\omega) = \sigma_i\}.$$

This is a modest extension of the langauge $\mathcal{L}_B(\Phi_\Gamma)$; all we have done is add a second batch of primitive propositions behaving very much the same way that the original formulas $play_i(\sigma_i)$ behave. One important difference between the two lies in how players consider them counterfactually, namely, in comparing expected utilities. Informally, players can evaluate what their utility would be if they were to play a different strategy, but *not* what their utility would be if they were to *intend* to play a different strategy.

In Section 5.2, we noted that our interpretation of gain-loss utility $n(t, p \,|\, s, q)$

entailed that $t = s$. Here we alter this interpretation: we assume instead that the reference level s is determined at a state ω by the player's *intention* at that state, rather than the actual strategy being played (which determines t). Accordingly, we define $u_C : \mathcal{S}(\mathcal{L}_B(\Phi_\Gamma^{int})) \to \mathbb{R}$ by

$$u_C(S) = |ref_C(S)|^{-1} \sum_{q \in ref_C(S)} u(t, p \,|\, s, q),$$

where $t = \rho_C(S)$, $p = \rho_R(S)$, and s is the unique element of Σ_C satisfying $int_C(s) \in S$.

We now consider a scenario where there is price certainty. Fix an intentional Γ-structure M and suppose that ω is a state at which C is certain that the shoes will be offered for price p. Suppose also that $\iota_C(\omega) = s$ and $s > p$. In other words, at state ω, C intends to buy the shoes.

A rational consumer, as always, seeks to maximize expected utility. Since she is uncertain about neither the price of the shoes nor her intention to buy them, her expected utility on playing $t \in T$ is given by $u(t, p \,|\, s, p)$. Let $t_L, t_H \in T$ be such that $t_L < p < t_H$. It is easy to calculate

$$u(t_L, p \,|\, s, p) = \eta p - \lambda \eta$$

and

$$u(t_H, p \,|\, s, p) = 1 - p;$$

therefore, a rational consumer will choose t_H rather than t_L just in case

$$p < \frac{1 + \lambda \eta}{1 + \eta}.$$

Thus, intending to buy makes it rational to buy even for some prices $p > 1$. In a situation where $s < p$, on the other hand, a similar calculation shows that a rational consumer will choose t_H over t_L only when

$$p < \frac{1 + \eta}{1 + \lambda \eta},$$

so intending not to buy makes is rational not to buy even for some prices $p < 1$. These findings duplicate those of KR.

6 Bayesian games

A *Bayesian game* is a model of strategic interaction among players whose preferences can depend on factors beyond the strategies they choose to play. These factors are often taken to be characteristics of the players themselves, such as whether they are industrious or lazy, how strong they are, or how they value certain objects. Such characteristics can be relevant in a variety of contexts: a job interview, a fight, an auction, etc.

To capture the factors beyond strategies, in a Bayesian game, we associate with each player a set of possible *types* of that player. These types can be thought of as encoding private information about the players. At a high level, we might imagine translating an $\mathcal{L}_B(\Phi_\Gamma)$-game into a Bayesian game by specifying, for each player i and situation S, player i's *type in* S to be the collection $t_i(S) = \{\varphi : B_i\varphi \in S\}$; in other words, player i's type is identified with the set of descriptions she is sure of. In a Bayesian game, utility depends on type, and in principle this kind of translation might allow one to transform situation-dependence into type-dependence, thereby capturing $\mathcal{L}_B(\Phi_\Gamma)$-games in the Bayesian framework. Are situations just a logical rendering of types?[9]

The sitauation is not quite so simple. There are standard notions of (Bayes-)Nash equilibrium in Bayesian games. There are standard conditions that guarantee the existence of a Bayesian-Nash equilibrium for a general class of Byaesian games. By of contrast, Proposition 4.2 shows that some quite simple language-based games do not have a Nash equilibrium. So, however we translate these language-based games to Bayseian games, the resulting Bayesian game must violate the conditions that guarantee the existence of an equilibrium.

This raises the obvious question of where exactly the intuition above goes wrong (or, at least, why it results in a Bayesian game that does not satisfy the conditions guaranteed to satisfy the existence of an equilibrium). It turns out that the real issue is whether or not types encode strategies. If they do not, we find the Bayesian formalism is simply not rich enough to accomodate language-based games, at least, not if the language can talk about strategies. On the other hand, if they do, then the standard arguments proving the existence of an equilibrium break down. Roughly speaking, the strategy that a player i should play at a type τ in equilibrium might not be the one that he actually does play, according to type τ.

These considerations lead us to associate two strategies with each player, an "intended" and an "actual" strategy. We can think of the intended strategy of player i at type τ as being the one encoded in his type, while the actual strategy is the one that he plays say in a given situation. As we show by example, this distinction between actual and intended types is quite useful.

In this section, we formalize some of the intuitions above, and do a preliminary investigation into these issues.

6.1 Definitions

A **Bayesian game** is a tuple $\mathcal{B} = (N, (\Sigma_i, T_i, p_i, u_i)_{i\in N})$ where

[9] We are indebted to Aviad Heifetz for suggesting this line of inquiry.

- $N = \{1, \dots, n\}$ is the set of *players*;
- Σ_i is the set of *strategies available to player i*;
- T_i is the measurable space of *types of player i*;
- $p_i : T_i \to \Delta(T_{-i})$ associates with each type t_i of player i a probability measure $p_i(t_i)$ on T_{-i} representing *player i's beliefs* about the types of her opponents;
- $u_i : \Sigma \times T \to \mathbb{R}$ is *player i's utility function.*

As we said, types can be thought of as encoding private information about the players.[10] This information is payoff-relevant in the sense that the utility functions depend on it as well as the actual strategies that are played. For example, the resolution of a battle between two armies may depend not only on what maneuvers they each perform (i.e. the strategies they employ), but also on how large or well-trained they were to begin with (i.e. their types). For a different kind of example: how happy one is with the results of an auction depends not only on who got what (determined by the bids that were placed, i.e. the strategies), but also on how the objects up for auction are valued (participants may value the same objects differently, which can be encoded by their types).

It is sometimes helpful to view a type $t_i \in T_i$ as determining the function $u_i(\cdot, t_i, \cdot) : \Sigma \times T_{-i} \to \mathbb{R}$, representing player i's preferences over the outcomes and types of her opponents. In other words, although t_i affects player i's utility, instead of thinking of player i as preferring to be one type rather than another, we think of her type as determining her preferences. However, this is a conceptual difference that plays little role in the formal developement.

Another key feature of the types formalism is that *types encode beliefs* via the functions p_i. It is often assumed that for all $t_i \in T_i$, the measure $p_i(t_i)$ is obtained from some fixed probability measure $\pi_i \in \Delta(T)$ by conditioning on t_i; in other words, each player's beliefs are obtained by conditioning her "prior beliefs" π_i on her own private information. When $\pi_1 = \pi_2 = \cdots = \pi_n$, we say that the players have a *common prior*; this condition is also frequently assumed in the literature.

Of course, a player's beliefs may not depend on her type in any essential way; for example, there is no reason why a commander's beliefs about the size of her opponent's army should depend on the size of her own army. But such a dependency *can* be encoded if desired: perhaps the commander of a very well-trained army tends to overestimate the discipline of her opponent's forces. Modeling beliefs about types is crucial for the analysis of many games. Consider once again an auction: several players place bids on several items, and a strategy for player i is identified with the collection of bids she places. As noted, different

[10] One can also introduce an extra player, "nature", whose type encodes information about "the world".

players may value the items up for auction differently; classically, we could simply encode this in the utility functions of the players. However, this leaves out an important aspect of the auction: the players may be uncertain about how their opponents value the items, and this information may be quite relevant to their own bidding strategies. Such a scenario, where the players are uncertain about the utility functions of their opponents, is a canonical example of the kind of strategic interaction that Bayesian games are designed to model. The types formalism captures this aspect of an auction simply and cleanly by encoding the players' valuations of the items in their types.

Of special interest for our purposes, the expressive power of the types formalism extends even further than this: because types encode beliefs, and utility depends on types, in principle Bayesian games can capture "psychological" effects in preferences, namely, preferences that depend on beliefs. For instance, one can define a Bayesian game in which a player's preferred strategy depends on whether or not her opponent is certain of her type. Given these considerations, it is natural to wonder to what extent language-based games are subsumed by the Bayesian framework.

6.2 Equilibrium and surprise

Part of the value of Bayesian games lies in the fact that a generalized notion of Nash equilibrium can be defined in this framework. A **Bayesian Nash equilibrium** of the Bayesian game $\mathcal{B}$ is a profile of *behaviour rules* $\beta_i : T_i \to \Delta(\Sigma_i)$ such that for each player i and each type t_i, the mixed strategy $\beta_i(t_i)$ maximizes player i's expected utility given the beliefs $p_i(t_i)$ and the behaviour rules β_{-i} of the other players. Note that the beliefs $p_i(t_i)$ that player i has about the *types* of her opponents yield beliefs about the *strategies* of her opponents via the behaviour rules β_{-i}. If we think of mixed strategies as representing conscious randomizations, we can think of this analogously to a classical Nash equilibrium, except here players choose mixtures that depend on their types, and rather than everyone knowing the mixture their opponents will use, everyone knows the mixtures that *each type* of their opponents will use. On the other hand, if we view a mixed strategy of player i as representing the common conjecture of her opponents about which (pure) strategy she will choose, then in a Bayesian Nash equilibrium, although the players may not have a common conjecture about their opponents' strategies, they *do* have a common conjecture about their opponents' strategies *given their types* (about which their beliefs may differ).

Every Bayesian game with finite strategy and type spaces admits a Bayesian Nash equilibrium [**?**]. By contrast, not every language-based game has a Nash equilibrium (Proposition 4.2). What explains this discrepancy?

The key impediment to the existence of a Nash equilibrium in the indignant altrusim game is the players' desire to surprise their opponents. Consider an arbitrary two-player game between Alice and Bob in which Bob prefers to play a strategy that Alice does not believe he will play (the exact interpretation of "does not believe" here—that is, whether it means "assigns low probability to", "assigns 0 probability to", or something else—is not important at the moment). Can a Bayesian game capture such a preference?

Speaking at a high level, Bob's utility function must be defined so that it assigns higher values to precisely those strategy-type profile pairs in which Bob's own strategy is unexpected with respect to the beliefs encoded by Alice's type. But this is easier said than done: types encodes beliefs *about types*, not about strategies, so Alice's type does not inherently specify her degree of belief about any of Bob's particular strategies. Though it is true that in the context of a Bayesian Nash equilibrium types induce beliefs about strategies (via the behaviour rules), this is far from sufficient for our purposes: we should be able to represent Bob's preference for surprise independently of any equilibrium constraints.

Thus, the discrepancy between equilibrium existence results in language-based versus Bayesian games is simply due to the fact that the latter is not expressive enough to represent the kinds of preferences that block equilibria in the former. Of course, this immediately begs the question: can we correct this deficiency? Types, in virtue of their abstract nature, are often conceived of as "catch-all" objects capable of encoding essentially anything that might be relevant to player preferences. However, while faith in the expressive power of the types formalism itself is not necessarily misplaced, in a Bayesian game enriching the type space can change the analysis in a fundamental way.

A natural enrichment is to encode strategies in types, so that each type t_i determines not only the preferences and beliefs of player i but also the strategy $s_i(t_i)$ that she is employing. In this case, of course, beliefs about types yield beliefs about strategies, circumventing the obstacle raised above and allowing us to model players who wish to surprise their opponents. But now the definition of Bayesian Nash equilibrium is muddied, since a behaviour rule also associates a strategy with a type. What is the relationship between $s_i(t_i)$ and $\beta_i(t_i)$?

6.3 Intended strategies

A **Bayesian game with intentions** $\mathcal{I}$ is a Bayesian game $\mathcal{B}$ equipped with additional functions $s_i : T_i \to \Sigma_i$ associating with each type t_i of player i an *intended strategy* $s_i(t_i)$. Intuitively, we might think of $s_i(t_i)$ as the strategy that a player of type t_i is *planning* to play (though may ultimately decide not to), or perhaps as the "default" strategy for that type. Regardless, as observed above, the functions s_i allow us to derive beliefs about strategies from beliefs about types, *even out of equilibrium.* In particular, we can represent Bob's preference to surprise Alice by defining Bob's utility function as follows:

$$u_B(\sigma, t) = \begin{cases} 1 & \text{if } p_A(t_A)(s_B^{-1}(\sigma_B)) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, Bob is happiest when the strategy σ_B that he is *actually* playing is such that Alice's type t_A considers it to be probability zero (remember that $p_A(t_A)$ is a measure on types, which is why we apply it to $s_B^{-1}(\sigma_B)$, i.e. the set of types for Bob whose intended strategy is σ_B).

A **Nash equilibrium of $\mathcal{I}$** is a profile of behaviour rules $\beta_i : T_i \to \Delta(\Sigma_i)$ that is a Bayesian Nash equilibrium of $\mathcal{B}$ and such that, for each $i \in N$ and $t_i \in T_i$, $s_i(t_i) \in \mathit{supp}(\beta_i(t_i))$, where *supp* denotes the support of the measure (i.e. the set of all strategies of positive probability). Said differently, a Nash equilibrium of $\mathcal{I}$ is just a Bayesian Nash equilibrium in which the common conjecture about what strategy a given type t_i is playing assigns positive probability to that type's intended strategy $s_i(t_i)$.

Not every Bayesian game with intentions admits a Nash equilibrium. For instance, let $\mathcal{I}$ be such that $\Sigma_A = \{*\}$, $\Sigma_B = \{\sigma_B, \sigma'_B\}$, $T_A = \{t_A, t'_A\}$, $T_B = \{t_B, t'_B\}$, $p_A(t_A)$ is the point-mass measure concentrated on t_B, $p_A(t'_A)$ is the point-mass measure concentrated on t'_B, $p_B(t_B)$ is the point-mass measure concentrated on t_A, $p_B(t'_B)$ is the point-mass measure concentrated on t'_A, $s_B(t_B) = \sigma_B$, $s_B(t'_B) = \sigma'_B$, and u_B is defined as above. Then we have

$$u_B(*, \sigma_B, t_A, t_B) = 0 < 1 = u_B(*, \sigma'_B, t_A, t_B),$$

and it follows that if β_B is part of a Bayesian Nash equilibrium, then $\beta_B(t_B)$ must be the pure strategy σ'_B (or else Bob's type t_B is not maximizing expected utility). On the other hand, since $s_B(t_B) = \sigma_B$, such a behaviour rule β_B cannot be part of a Nash equilibrium for $\mathcal{I}$. Thus, $\mathcal{I}$ has no Nash equilibrium.

It should not be terribly surprising that Bayesian games with intentions do not always have Nash equilibria: such a framework can capture the desire to surprise, and we have seen (Example 3.4 and Section 4.2) that this kind of preference is intuitively at odds with the notion of Nash equilibrium. But aside from modeling the desire to surprise, incorporating "intention" into games is of interest in its own right. In some cases a player may have a "default" course of action, a strategy that is distinguised from the others—perhaps deviating from it incurs a small cost [REF?]. We also saw in Section 5.3 that a notion of intention is useful for capturing KR's *personal equilibrium* solution concept. More generally, representing intention can be important in contexts where the distinction between present plans and future actions is of import. Example 3.8, for instance, might be viewed as analyzing certain kinds of procrastination as the overvaluing of intended actions. Further research in this area seems promising.

7 Conclusion

Language-based games generalize classical games by replacing outcomes with situations as the objects of preference. The underlying language determines the extent of this generalization. We saw, for example, that situations correspond exactly to outcomes with the right choice of underlying language.

Language-based games also generalize psychological games. Broadly speaking, psychological game theory is concerned with cases where the beliefs of the players are relevant to their preferences; in this informal sense, any language-based game where the underlying language expresses the beliefs of the players in some way is an instance of a psychological game. More formally, however, psychological games allow the players' utility functions to depend on beliefs in a continuous way, so to fully subsume this theory requires an underlying language rich enough to do the same, such as the language $\mathcal{L}_B^{[0,1]}(\Phi_\Gamma)$ defined in Example 3.5.

In this paper, we have focussed primarily on the language $\mathcal{L}_B(\Phi_\Gamma)$, with occasional forays into richer representations of belief (Example 3.5) and notions of "intention" (Example 3.8 and Section 5.3). Further investigation into the role of intention in decision making, and particularly its connection to procrastination, seems promising. Moreover, there are several other natural extensions of the underlying language worthy of study. Temporal logics [14], for example, offer an appealing avenue for extending language-based games to a dynamic setting. Logics of awareness [12], on the other hand, offer a potential route by which to incorporate uncertainty *about the underlying language* into the game. By a slight generalization of the current framework, we can assign different players distinct underlying languages at each state, which allows each player i to be uncertain about what language his opponents' preferences are defined over. Indeed, reasoning about how opponents conceptualize the world insofar as their preferences are concerned is quite relevant to a variety of strategic interactions; it is, for instance, presumably what is at play when retailers set prices like \$2.99 rather than \$3, exploiting the "rounding error" consumers typically make (Example 3.1).

Finally, as we have emphasized (Examples 3.1, 3.2, and 3.3), coarseness can be a powerful tool in the resolution of apparent paradoxes of human decision making. Coarseness in this sense can be viewed as an implementation of *bounded rationality*: players do not represent the world in all its gory detail, but rather, they systematically collapse certain distinctions by subsuming them under the same description. While the notion of bounded rationality is certainly not new, studying it through the lens of language provides an intuitive and simple mechanism with which a wide variety of decision problems can be analyzed. Moreover, the technical advantages of this implementation are apparent in, for example, equilibrium analyses that do not depend on continuity of the utility functions, such as Corollary 4.9.

A Proofs

Proposition 4.1: For each $i \in N$, the map $\vec{\rho}_i : \mathcal{S} \to \Sigma_i \times \mathcal{S}_{-i}$ defined by $\vec{\rho}_i(S) = (\rho_i(S), \rho_{-i}(S))$ is a bijection.

Proof:

Fix $i \in N$. To show that $\vec{\rho_i}$ is surjective, it suffices to show that given a situation S and a strategy $\sigma_i \in \Sigma_i$, there exists a situation S' such that $\rho_{-i}(S') = \rho_{-i}(S)$, and $\rho_i(S') = \sigma_i$. Let M be a Γ-structure with a state ω such that $(M, \omega) \models S$; we will find the desired S' by constructing a new Γ-structure M' in which, intuitively, there is a state that is just like ω except that player i is playing σ_i.

Define $\tilde{\Omega} = \Omega \times \{1, 2\}$. For each $\omega' \in \Omega$, set $\tilde{s}_i(\omega', 2) = \sigma_i$; for $(j, k) \neq (i, 2)$, define $\tilde{s}_j(\omega', k) = s_j(\omega')$. Similarly, for each $\omega' \in \Omega$, let $\tilde{\mathcal{PR}}_i(\omega', 2)$ be the measure on $\Omega \times \{2\}$ induced by $\mathcal{PR}_i(\omega')$ by the natural correspondence; for $(j, k) \neq (i, 2)$, let $\tilde{\mathcal{PR}}_j(\omega', k)$ be the measure on $\Omega \times \{1\}$ induced by $\mathcal{PR}_j(\omega')$ by the natural correspondence.

It is not hard to check that $\tilde{M} = (\tilde{\Omega}, (\tilde{s}_i)_{i \in N}, (\tilde{\mathcal{PR}}_i)_{i \in N})$ is a Γ-structure. Intuitively, it contains a "copy" of M in the component corresponding to $\Omega \times \{1\}$, while the component corresponding to $\Omega \times \{2\}$ is just like M except that player i is playing σ_i at all states. It is now easy to show by induction on the structure of formulas that if φ is i-independent, then $(M, \omega') \models \varphi$ iff $(\tilde{M}, (\omega', 2)) \models \varphi$ and that, for all formulas φ, $(M, \omega') \models \varphi$ iff $(\tilde{M}, (\omega', 1)) \models \varphi$. Taking $S' = S(\omega, 2)$ therefore yields the desired result.

Now we show that $\vec{\rho_i}$ is injective; for this it suffices to show that if $S, S' \in \mathcal{S}$ are distinct situations with $\rho_{-i}(S) = \rho_{-i}(S')$, then $\rho_i(S) \neq \rho_i(S')$. This follows easily from the following claim, which we prove by structural induction: for all formulas φ, if φ is not i-independent, then the subsets $X_\varphi, Y_\varphi \subseteq \Sigma_i$ defined by

$$\begin{aligned} X_\varphi &:= \{\sigma_i : \varphi \not\models \neg play_i(\sigma_i)\} \\ Y_\varphi &:= \{\sigma_i : \neg\varphi \not\models \neg play_i(\sigma_i)\} \end{aligned}$$

are disjoint. In other words, whenever φ is not i-independent, the set of strategies for player i compatible with φ is disjoint from the set of strategies compatible with $\neg\varphi$. Now if S and S' are distinct situations with $\rho_{-i}(S) = \rho_{-i}(S')$, then they must differ on some formula φ that is not i-independent. Therefore, by the above, we can conclude that $\rho_i(S) \neq \rho_i(S')$, proving injectivity.

For the base case, suppose $\varphi = play_j(\sigma_j)$. If $j \neq i$ then this formula is i-independent and we are done; otherwise, it is easy to see that $X_{play_i(\sigma_i)} = \{\sigma_i\}$ and $Y_{play_i(\sigma_i)} = \Sigma_i \setminus \{\sigma_i\}$, which are certainly disjoint. The inductive step for negation follows easily from the observation that $X_{\neg\varphi} = Y_\varphi$ and $Y_{\neg\varphi} = X_\varphi$. If the result holds for φ and ψ, then since

$$X_{\varphi \wedge \psi} \subseteq X_\varphi \cap X_\psi$$

and

$$Y_{\varphi \wedge \psi} = X_{\neg\varphi \vee \neg\psi} = X_{\neg\varphi} \cup X_{\neg\psi} = Y_\varphi \cup Y_\psi,$$

it follows that $X_{\varphi \wedge \psi}$ and $Y_{\varphi \wedge \psi}$ are disjoint, which establishes the inductive step for conjunction. The inductive step for B_j, $j \neq i$, is trivial since the resulting formula is i-independent. Finally, note that whenever $\varphi \models \neg play_i(\sigma_i)$, we also have $B_i\varphi \models B_i\neg play_i(\sigma_i)$, and so $B_i\varphi \models \neg play_i(\sigma_i)$. It follows that $X_{B_i\varphi} \subseteq X_\varphi$, and similarly $Y_{B_i\varphi} \subseteq Y_\varphi$, from which disjointness follows. ▮

Lemma 4.5: *$EB^k(RAT)$ is satisfiable for all $k \in \mathbb{N}$.*

Proof: The idea is to construct a Γ-structure that is particularly well-behaved with respect to alterations of its strategy function; this will allow us to modify a given strategy function in such a way as to ensure that the players are rational at certain states.

Let T be the set of all finite words on the alphabet N (the set of players), excluding those words in which any letter appears consecutively:

$$T \coloneqq \{w \in N^* \; : \; (\forall i < |w| - 1)[w(i) \neq w(i+1)]\}.$$

Thus T can be viewed as a tree whose root node λ (the empty word) has $n = |N|$ children, while every other node has $n - 1$ children (one for each letter in N aside from the last letter of the current node). This will be our state space.

Given any nonempty word w, let $\ell(w) \coloneqq w(|w| - 1)$, the last letter in w. Define $\mathcal{PR}_i(w) \coloneqq \delta_{succ_i(w)}$, the point-mass probability measure concentrated on $succ_i(w) \in T$, where (taking $w \cdot i$ to be the result of appending i the end of w)

$$succ_i(w) \coloneqq \begin{cases} i & \text{if } w = \lambda \\ w \cdot i & \text{if } \ell(w) \neq i \\ w & \text{otherwise.} \end{cases}$$

It is easy to see that the *frame* (i.e., Γ-structure without the strategy function $s = (s_i)_{i \in N}$) $F = (T, \mathcal{PR}_1, \ldots, \mathcal{PR}_n)$ satisfies conditions (P1) through (P3); in particular, (P3) follows from the observation that $succ_i$ is idempotent.

Note that, given strategy functions s, (F, s) is a Γ-structure. Our goal is to define strategy functions s on T in such a way as to ensure that $((F, s), \lambda) \models EB^k(RAT)$. Note that $((F, s), \lambda) \models EB^k(RAT)$ just in case $((F, s), w) \models RAT$ for every word w with $|w| \leq k$. We prove that this can be arranged by induction on k. More precisely, we prove the following statement by induction on k:

For all $k \in \mathbb{N}$ and $s : T \to \Sigma$, there exists an $s' : T \to \Sigma$ such that

- *(i) for all w with $|w| > k + 1$, $s'(w) = s(w)$;*
- *(ii) for all w with $|w| = k + 1$ and all $i \neq \ell(w)$, $s'_i(w) = s_i(w)$;*
- *(iii) for all w with $|w| \leq k$, $((F, s'), w) \models RAT$.*

The additional assumptions (i) and (ii) in this statement allow us to apply the inductive hypothesis without fear of causing RAT to fail at nodes where we previously established that it held.

For the base case $k = 0$, let s be a given strategy function. For each $i \in N$, let $\sigma_i \in BR_i(\lambda)$ (recall that the best response function depends on the state). Define $s'(\lambda) := (\sigma_1, \ldots, \sigma_n)$. In order to satisfy (P4), we must also insist that for each $j \in N$, $s'_j(\lambda \cdot j) = \sigma_j$. Otherwise, let s' agree with s. Then it is easy to see that $((F, s'), \lambda) \models RAT$, since we have altered each player's strategy at λ so as to ensure its rationality. It is also clear from construction that condition (i) is satisfied, and moreover for each $j \in N$ and each $i \neq j$ we have $s'_i(\lambda \cdot j) = s_i(\lambda \cdot j)$, so condition (ii) is satisfied as well. This completes the proof for the base case.

For the inductive step, assume the statement holds for k, and let s be a given strategy function. Roughly speaking, we first modify s so that RAT holds at all words of length $k+1$, and then appeal to the inductive hypothesis to further modify the strategy function so that RAT holds at all words of length $\leq k$. For each word w of length $k+1$, and for each $i \neq \ell(w)$, choose $\sigma_i \in BR_i(w)$ and redefine s so that player i is playing σ_i at w and at $w \cdot i$. Call the resulting strategy function s'. Similarly to the base case, it is easy to see that for each w of length $k+1$ and $i \neq \ell(w)$, we have $((F, s'), w) \models RAT_i$.

Applying the inductive hypothesis to s', we obtain a new strategy function s'' such that for all w with $|w| \leq k$, $((F, s''), w) \models RAT$. It follows that for each word w of length k and each $i \in N$, $((F, s''), succ_i(w)) \models RAT_i$, since $\mathcal{PR}_j(w) = \mathcal{PR}_j(succ_i(w))$. Moreover, from conditions (i) and (ii) we can deduce that the property we arranged above for words w of length $k+1$, namely that $((F, s'), w) \models RAT_i$ for each $i \neq \ell(w)$, is preserved when we switch to the strategy function s''. Moreover, if w has length $k+1$ and $i = \ell(w)$, so that $w = w' \cdot i$, then $((F, s''), w) \models RAT_i$ since $((F, s''), w') \models RAT_i$ by the inductive hypothesis. Putting these facts together, we see that for each word w of length $k+1$, we have $((F, s''), w) \models RAT$. Thus for all w with $|w| \leq k+1$ we have $((F, s''), w) \models RAT$; conditions (i) and (ii) are straightforward to verify. This completes the induction. ∎

Theorem 4.7 *(CR) implies that rationalizable strategies exist.*

Proof: Assuming (CR), we define an iterative deletion procedure on situations. First, let

$$\mathcal{R} = \{S \in \mathcal{S} \,:\, S \not\models \neg RAT\}.$$

Thus, $S \in \mathcal{R}$ precisely when S is compatible with rationality; that is, when $S \cup \{RAT\}$ is satisfiable. Condition (CR) has a particularly nice topological formulation in terms of $\mathcal{R}$.

Lemma A.1: *(CR) holds if and only if $\mathcal{R}$ is closed in $\mathcal{S}$.*

Proof: Suppose $S \notin \mathcal{R}$. Then, by definition, $S \models \neg RAT$, so (CR) guarantees that there is some finite subset $F \subset S$ such that $F \models \neg RAT$. In fact, since S

is maximal, it easy to see that the formula

$$\varphi_S \coloneqq \bigwedge_{\psi \in F} \psi$$

is itself an element of S, so without loss of generality we can replace the set F with the single formula φ_S. U_{φ_S} is open, by definition. Moreover, $U_{\varphi_S} \cap \mathcal{R} = \emptyset$, since any set $S' \in U_{\varphi_S}$ contains φ_S, and therefore entails $\neg RAT$. Since $S \in U_{\varphi_S}$, this establishes that $\mathcal{R}$ is closed.

Conversely, suppose that $\mathcal{R}$ is closed in $\mathcal{S}$, and let $S \in \mathcal{S}$ be such that $S \models \neg RAT$. Then $S \notin \mathcal{R}$, so there is some basic open set U_φ such that $S \in U_\varphi$ and $U_\varphi \cap \mathcal{R} = \emptyset$. Thus $\varphi \in S$, and any situation that contains φ must entail $\neg RAT$, from which it follows that $\varphi \models \neg RAT$. ∎

Having defined those situations not compatible with rationality, we next define the iterative portion of the deletion procedure, designed to yield all and only those situations compatible with common belief of rationality.

By Lemma A.1, $\mathcal{R}$ is closed, so we can express its complement as a union of basic open sets: let $I \subset \mathcal{L}_B(\Phi_\Gamma)$ be such that

$$\mathcal{R} = \mathcal{S} - \bigcup_{\varphi \in I} U_\varphi.$$

Note that, by definition, S is not compatible with rationality just in case S contains some formula in I. Roughly speaking, we can think of I as an exhaustive list of the ways in which rationality might fail. We therefore define

$$\mathcal{R}^{(1)} = \{S \in \mathcal{R} \, : \, (\forall i \in N)(\forall \varphi \in I)[B_i \neg \varphi \in S]\}.$$

Intuitively, $\mathcal{R}^{(1)}$ is the set of situations that are not only compatible with rationality, but in which each player *believes* that the situation is compatible with rationality (remember that "rationality" is being used here as a shorthand for "everyone is rational"). If we set

$$I^{(1)} = \{\widehat{B}_i \varphi \, : \, i \in N \text{ and } \varphi \in I\},$$

then we can express $\mathcal{R}^{(1)}$ more succintly as

$$\mathcal{R}^{(1)} = \mathcal{R} - \bigcup_{\psi \in I^{(1)}} U_\psi.$$

This also makes it clear that $\mathcal{R}^{(1)}$ is closed in $\mathcal{S}$. More generally, let $I^{(0)} = I$ and $\mathcal{R}^{(0)} = \mathcal{R}$; for each $k \geq 1$, set

$$I^{(k)} = \{\widehat{B}_i \varphi \, : \, i \in N \text{ and } \varphi \in I^{(k-1)}\},$$

and define

$$\mathcal{R}^{(k)} = \mathcal{R}^{(k-1)} - \bigcup_{\psi \in I^{(k)}} U_\psi.$$

It is straightforward to check that this definition agrees with our original definition of $\mathcal{R}^{(1)}$ and $I^{(1)}$. Moreover, observe that

$$\mathcal{R}^{(0)} \supseteq \mathcal{R}^{(1)} \supseteq \mathcal{R}^{(2)} \supseteq \cdots$$

is a nested, decreasing sequence of closed subsets of $\mathcal{S}$. Since $\mathcal{S}$ is compact, a collection of closed sets with the finite intersection property[11] has nonempty intersection.

Lemma A.2: *For all $k \in \mathbb{N}$ and $S \in \mathcal{S}$, if $S \cup \{EB^k(RAT)\}$ is satisfiable, then $S \in \mathcal{R}^{(k)}$.*

Proof: The proof proceeds by induction on k. For the base case $k = 0$, we must show that if $S \cup \{RAT\}$ is satisfiable, then $S \in \mathcal{R}$, which is precisely the definition of $\mathcal{R}$.

Now suppose inductively that the statement holds for $k - 1$, and let $S \in \mathcal{S}(\mathcal{L}_B(\Phi_\Gamma))$ be such that $S \cup \{EB^k(RAT)\}$ is satisfiable. Then $S \cup \{EB^{k-1}(RAT)\}$ is also satisfiable, so by the inductive hypothesis we know that $S \in \mathcal{R}^{(k-1)}$. Therefore, by definition of $\mathcal{R}^{(k)}$, the only way we could have $S \notin \mathcal{R}^{(k)}$ is if $\widehat{B}_i \varphi \in S$ for some $i \in N$ and $\varphi \in I^{(k-1)}$. Suppose for contradiction that this is so.

By assumption, there is some Γ-structure $M = (\Omega, (s_i)_{i \in N}, (\mathcal{PR}_i)_{i \in N})$ and some $\omega \in \Omega$ such that $\omega \models S \cup \{EB^k(RAT)\}$. Furthermore, since $\widehat{B}_i \varphi \in S$, we must have $\mathcal{PR}_i(\omega)(\llbracket \varphi \rrbracket_M) > 0$. However, for a state $\omega' \in \llbracket \varphi \rrbracket_M$, by definition, $S(\omega') \notin \mathcal{R}^{(k-1)}$ (since $\varphi \in I^{(k-1)}$). By the inductive hypothesis, it follows that $S(\omega') \cup \{EB^{k-1}RAT\}$ is not satisfiable, so in particular $\omega' \not\models EB^{k-1}RAT$. We have therefore shown that $\llbracket \varphi \rrbracket_M \cap \llbracket EB^{k-1}RAT \rrbracket_M = \emptyset$, from which we can conclude that $\mathcal{PR}_i(\omega)(\llbracket EB^{k-1}RAT \rrbracket_M) < 1$, contradicting the fact that $\omega \models EB^k RAT$. ∎

In light of Lemma 4.5, Lemma A.2 implies that for each $k \in \mathbb{N}$, $\mathcal{R}^{(k)} \neq \emptyset$. Therefore the collection $\{\mathcal{R}^{(k)} : k \in \mathbb{N}\}$ does indeed have the finite intersection property, hence

$$\mathcal{R}^\infty := \bigcap_{k=0}^{\infty} \mathcal{R}^{(k)} \neq \emptyset.$$

The following lemma therefore clinches the main result.

Lemma A.3: *$S \in \mathcal{R}^\infty$ if and only if $S \cup \{CB(RAT)\}$ is satisfiable.*

Proof: One direction is easy: if $S \cup \{CB(RAT)\}$ is satisfiable, then for every $k \in \mathbb{N}$ we know that $S \cup \{EB^k(RAT)\}$ is satisfiable. Lemma A.2 then guarantees

[11] Recall that a collection of sets has the *finite intersection property* just in case every finite subcollection has nonempty intersection.

that

$$S \in \bigcap_{k \in \mathbb{N}} \mathcal{R}^{(k)} = \mathcal{R}^\infty,$$

as desired.

Now we prove the converse. Suppose that $\mathcal{R}^\infty \neq \emptyset$; for each $S \in \mathcal{R}^\infty$, let $M^S = (\Omega^S, (s_i^S)_{i \in N}, (\mathcal{PR}^S)_{i \in N})$ be a Γ structure with a distinguished state $\omega^S \in \Omega^S$ such that $\omega^S \models S \cup \{RAT\}$. This is always possible because $S \in \mathcal{R}^\infty \subseteq \mathcal{R}$. Let

$$\tilde{\Omega} = \bigsqcup_{S \in \mathcal{R}^\infty} (\Omega^S \times N),$$

and equip this set with the σ-algebra of measurable sets generated by all sets of the form

$$\bigsqcup_{S \in \mathcal{R}^\infty,\, i \in N} E_{S,i},$$

where $E_{S,i}$ is a measurable subset of $\Omega^S \times \{i\}$. For $i \in N$ and $(\omega, k) \in \Omega^S \times N$, set $\tilde{s}_i(\omega, k) = s_i^S(\omega)$ and define

$$\tilde{\mathcal{PR}}_i(\omega, k) = \begin{cases} \mathcal{PR}_i^{S'}(\omega^{S'}) \restriction \Omega^{S'} \times \{i\} & \text{if } i \neq k \text{ and } S' := S(M^S, \omega) \in \mathcal{R}^\infty \\ \mathcal{PR}_i^S(\omega) \restriction \Omega^S \times \{i\} & \text{otherwise,} \end{cases}$$

where the symbol $\restriction$ in the above is simply used to indicate that the probability measure given on the left is to be interpreted in the set given on the right via the natural correspondence (so, for example, though $\mathcal{PR}_i^S(\omega)$ is technically defined over Ω^S, we can interpret it instead as being defined over $\Omega^S \times \{i\}$ since this space is isomorphic to Ω^S).

It is straightforward (if tedious) to show that $\tilde{M} := (\tilde{\Omega}, (\tilde{s}_i)_{i \in N}, (\tilde{\mathcal{PR}}_i)_{i \in N})$ is a Γ-structure, and moreover that it has the following property: for all formulas $\varphi \in \mathcal{L}_B(\Phi_\Gamma)$ and every $(\omega, k) \in \Omega^S \times N$,

$$(\tilde{M}, (\omega, k)) \models \varphi \text{ iff } (M^S, \omega) \models \varphi.$$

In particular, for all $k \in N$ and $S \in \mathcal{R}^\infty$, $(\tilde{M}, (\omega^S, k)) \models S$. Thus, if we show that $(\tilde{M}, (\omega^S, k)) \models CB(RAT)$ we will be done. For this it suffices to prove that $(\tilde{M}, (\omega^S, k)) \models EB^m(RAT)$ for all $m \in \mathbb{N}$, which follows by induction, employing the following crucial fact about $\tilde{M}$: for every $S \in \mathcal{R}^\infty$, $k \in N$, and $i \in N$,

$$\tilde{\mathcal{PR}}_i(\omega^S, k)(\{(\omega', k') \in \tilde{\Omega} \ : \ S(\tilde{M}, (\omega', k')) \in \mathcal{R}^\infty)\}) = 1.$$

This, in turn, is a consequence of the fact that $S(\tilde{M}, (\omega^S, k)) \in \mathcal{R}^\infty$, and therefore $\tilde{\mathcal{PR}}_i(\omega^S, k)$ assigns probability 0 to each of the (countably many) formulas in $I^\infty := \bigcup I^{(m)}$ which witness a situation not being in $\mathcal{R}^\infty$. ∎

Since $\mathcal{R}^\infty$ is nonempty, by Lemma A.3 there is some situation $S \in \mathcal{S}$ such that $S \cup \{CB(RAT)\}$ is satisfiable. Thus, the strategy profile $(\rho_1(S), \ldots, \rho_n(S)) \in \Sigma$ is rationalizable, as desired. ∎

B Acknowledgements

Bjorndahl is supported in part by NSF grants IIS-0812045, CCF-1214844, DMS-0852811, and DMS-1161175, and ARO grant W911NF-09-1-0281. Halpern is supported in part by NSF grants IIS-0812045, IIS-0911036, and CCF-1214844, by AFOSR grant FA9550-08-1-0266, by the Multidisciplinary University Research Initiative (MURI) program administered by the AFOSR under grant FA9550-12-1-0040, and by ARO grant W911NF-09-1-0281. Pass is supported in part by an Alfred P. Sloan Fellowship, a Microsoft Research Faculty Fellowship, NSF Awards CNS-1217821 and CCF-1214844, NSF CAREER Award CCF-0746990, AFOSR YIP Award FA9550-10-1-0093, and DARPA and AFRL under contract FA8750-11-2-0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

References

[1] M. Allais. Le comportement de l'homme rationel devant le risque: critique de l'école Americaine. *Econometrica*, 21:503–546, 1953.

[2] R. J. Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55:1–18, 1987.

[3] R. J. Aumann and A. Brandenburger. Epistemic conditions for Nash equilibrium. *Econometrica*, 63(5):1161–1180, 1995.

[4] P. Battigalli and M. Dufwenberg. Dynamic psychological games. *Journal of Economic Theory*, 144:1–35, 2009.

[5] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic.* Cambridge Tracts in Theoretical Computer Science, No. 53. Cambridge University Press, Cambridge, U.K., 2001.

[6] A. Brandenburger and E. Dekel. Rationalizability and correlated equilibria. *Econometrica*, 55:1391–1402, 1987.

[7] I. Brocas, J. D. Carrillo, and M. Dwatripont. Commitment devices under self-control problems: an overview. In I. Brocas and J. D. Carrillo, editors, *The Psychology of Economic Decisions: Volume II: Reasons and Choices*, pages 49–67. Oxford University Press, Oxford, UK, 2004.

[8] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge.* MIT Press, Cambridge, Mass., 1995. A slightly revised paperback version was published in 2003.

[9] J. Geanakoplos, D. Pearce, and E. Stacchetti. Psychological games and sequential rationality. *Games and Economic Behavior*, 1(1):60–80, 1989.

[10] J. Y. Halpern. *Reasoning About Uncertainty.* MIT Press, Cambridge, Mass., 2003.

[11] J. Y. Halpern and R. Pass. Game theory with translucent players. In *Proceedings of the Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 216–221, 2013.

[12] J. Y. Halpern and L. C. Rêgo. Reasoning about knowledge of unawareness revisited. *Mathematical Social Sciences*, 66(2):73–84, 2013.

[13] B. Kőszegi and M. Rabin. A model of reference-dependent preferences. *The Quarterly Journal of Economics*, CXXI:1133–1165, 2006.

[14] Z. Manna and A. Pnueli. *The Temporal Logic of Reactive and Concurrent Systems: Specification*. Springer-Verlag, Berlin/New York, 1992.

[15] C. F. Manski and F. Molinari. Rounding probabilistic expectations in surveys. *Journal of Business and Economic Statistics*, 28(4):219–231, 2010.

[16] B. J. McNeil, S. J. Pauker, H. C. Sox Jr., and A. Tversky. On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, 306:1259–1262, 1982.

[17] R. S. Moyer and T. K. Landauer. Time required for judgements of numerical inequality. *Nature*, 215:1519–1520, 1967.

[18] S. Mullainathan. Thinking through categories. Unpublished manuscript; available at www.haas.berkeley.edu/groups/finance/cat3.pdf, 2002.

[19] M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, Mass., 1994.

[20] F. Restle. Speed of adding and comparing numbers. *Journal of Experimental Psychology*, 83:274–278, 1978.

[21] B. Salcedo. Implementation without commitment in moral hazard environments. Working paper, 2013.

[22] T. Tan and S. Werlang. The Bayesian foundation of solution concepts of games. *Journal of Economic Theory*, 45(45):370–391, 1988.