Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

Language-Based Games (Extended Abstract)*

Adam Bjorndahl, Joseph Y. Halpern, Rafael Pass
Cornell University
Ithaca, New York, USA

Abstract

We introduce *language-based games*, a generalization of *psychological games* [Geanakoplos *et al.*, 1989] that can also capture *reference-dependent preferences* [Kőszegi and Rabin, 2006], which extend the domain of the utility function to *situations*, maximal consistent sets in some language. The role of the underlying language in this framework is thus particularly critical. Of special interest are languages that can express only *coarse* beliefs [Mullainathan, 2002]. Despite the expressive power of the 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.

1 Introduction

In a classical, normal-form game, an *outcome* is a tuple of strategies, one for each player. Players' preferences are formalized by utility functions defined on the set of all such outcomes. This framework thereby hard-codes the assumption that a player can prefer one state of the world to another only insofar as they differ in the outcome of the game.

Perhaps unsurprisingly, this model is too restrictive to account for a broad class of interactions that otherwise seem well-suited to a game-theoretic analysis. For example, 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 [1989] and expanded by Battigalli and Duwfenberg [2009], is an enrichment of the classical setting meant to capture such preferences and motivations. In a similar vein, *reference-dependent preferences* [Kőszegi and Rabin, 2006] 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.

In both of these approaches, the method of generalization takes the same basic form: the domain of a utility function is enlarged so as to include not only the outcomes of the game, but also the beliefs of the players. The resulting domain may be quite complex; for instance, in psychological game theory, since the goal is to model preferences that depend not only on beliefs about outcomes, but also beliefs about beliefs, beliefs about beliefs about beliefs, and so on, the domain includes infinite hierarchies of beliefs.

The model we present in this paper, though motivated in part by a desire to capture belief-dependent preferences, is geared towards a much more general goal. Besides being expressive enough to subsume existing systems such as those described above, it establishes a general framework for modeling players with richer preferences. Moreover, it is equally capable of representing *impoverished* preferences, a canonical example of which are so-called "coarse beliefs" or "categorical thinking" [Mullainathan, 2002]. Coarse beliefs (beliefs that take only a small number of possible probability values) often seem more natural than fine-grained (continuous) beliefs when it comes to modeling human preferences. As we show by example, utilities defined over coarse beliefs provide a natural way of capturing some otherwise puzzling behavior.

The core idea is that a player's preferences are expressed in some language. A player cannot place utility on something that she cannot express. The language may include beliefs (which allows us to capture psychological games), formulas that talk about both expected and actual outcomes (which allows us to capture reference-dependent preferences), and it may be quite restricted (which allows us to capture coarse beliefs). Players' preferences can be expressed in a simple and natural manner, narrowing the divide between intuition and formalism. As a preliminary illustration of some of these points, 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.

*The paper on which this extended abstract is based was nominated by the program chair as one of the best papers of the Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK), 2013 [Bjorndahl, Halpern, and Pass, 2013]. The full version is available at http://www.cs.cornell.edu/home/halpers/papers/lbg.pdf.

We might summarize this scenario by Table 1. In this table,

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

Table 1: The surprise proposal.

we denote Bob's two strategies, proposing and not proposing, as 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.

Granted, 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 there is good reason to think (see [Mullainathan, 2002]) that an accurate model of her expectations stops here, with some small *finite* 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:

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

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

We view Table 1 as *defining* Bob's utility. To coax an actual utility 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, these include 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 are calling a "situation", and 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 by using more than two categories; we spell out this straightforward generalization in the full paper. Indeed, we could allow a representation that permits continuous probabilities, as has been done in the literature. However, we will see that an "all-or-nothing" representation of belief is enough to capture some interesting and complex games. ∎

The central concept we develop in this paper is that of a *language-based game*, where utility is defined not on outcomes or the Cartesian product of outcomes with some other domain, but on *situations*. As noted, a situation can be conceptualized as a collection of formulas about the game (technically, a maximal consistent set of formulas) in an appropriate language, intuitively, the language that characterizes what the player's preferences depend on. Succinctly a player can prefer one state of the world to another if and only if she can *describe* the difference between the two, where "describe" here means "express in the underlying language".

Language-based games are thus parametrized by the underlying language: changing the language changes the game. The power and versatility of our approach derives in large part from this dependence. Consider, for example, an underlying language that contains only terms refering to players' strategies. With this language, players' preferences can depend only on the outcome of the game, as is the case classically. Thus, classical game theory can be viewed as a special case of the language-based approach of this paper (see Sections 2.1 and 2.2 for details).

Enriching the underlying language allows for an expansion and refinement of players' preferences; in this manner we are able to subsume, for example, work on psychological game theory and reference-dependent preferences, in addition to providing some uniformity to the project of defining new and further expansions of the classical base. By contrast, restricting the underlying language coarsens the domain of player preference; this provides a framework for modeling phenomena like coarse beliefs. A combination of these two approaches yields a theory of belief-dependent preferences incorporating coarse beliefs.

We give a few examples here of how the framework can be used; we encourage the reader to consult the full paper for more details. We make three major contributions. First, as noted, our system is easy to use in the sense that players' preferences are represented with a simple and uncluttered formalism; complex psychological phenomena can thus be captured in a direct and intuitive manner. Second, we provide a formal game-theoretic representation of coarse beliefs, and in so doing, expose 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 effects that have heretofore been modeled only in a continuous framework. The full paper provides several examples that illustrate these points. Third, we provide novel equilibrium analyses that do not depend on the continuity of the expected utility function as do Geanakoplos et al. [1989]. Note that such continuity assumptions are at odds with our use of coarse beliefs. In particular, our main theorem demonstrates that if the underlying language satisfies certain natural "compactness" assumptions, then every game over this language has rationalizable strategies; see the full paper for a formalization of this theorem. In contrast, even under these

assumptions about the language, not every game has a Nash equilibrium (see Example 2.1).

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 $\{1, \ldots, n\}$, and Σ_i is the set of *strategies available to player* i. Following standard notation, we set $\Sigma = \prod_{i \in N} \Sigma_i$ and $\Sigma_{-i} = \prod_{j \neq i} \Sigma_j$. 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 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: {"It's raining"} and {"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 while still being consistent in a sense 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

Given a set Φ of primitive propositions, let $\mathcal{L}(\Phi)$ denote the propostional logic based on 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 $play_1(\sigma_1) \wedge \cdots \wedge play_n(\sigma_n)$.

As usual, a set of formulas F is said to be *satisfiable (with respect to a set $\mathcal{M}$ of admissible models)* if there is some model in $\mathcal{M}$ in which every formula of F is true. An $\mathcal{L}(\Phi)$-*situation* is then defined to be a *maximal* satisfiable set of formulas (with respect to the admissible models of $\mathcal{L}(\Phi)$): that is, a satisfiable set with no proper superset that is also satisfiable. In the game-theoretic setting, an admissible model is a valuation that, for each player i, makes exactly one of the formulas $play_i(\sigma)$ true (since each player can choose exactly one strategy). Situations correspond to admissible models: a situation just consists of all the formulas true in some admissible model. Let $\mathcal{S}(\mathcal{L}(\Phi))$ denote the set of $\mathcal{L}(\Phi)$-situations. It is not difficult to see that $\mathcal{S}(\mathcal{L}(\Phi_\Gamma))$ can be identified with the set Σ of outcomes.

Having illustrated some of the principle concepts of our approach in the context of propositional logic, we now present the definitions in complete generality. Let $\mathcal{L}$ be a language with an associated semantics, that is, a set of admissible models providing a notion of truth. We often use the term "language" to refer to a set of well-formed formulas together with a set of admissible models (this is sometimes called a "logic"). An **$\mathcal{L}$-situation** is a maximal satisfiable set of formulas from $\mathcal{L}$. Denote by $\mathcal{S}(\mathcal{L})$ the set of $\mathcal{L}$-situations. A game form Γ is extended to an **$\mathcal{L}$-game** by adding utility functions $u_i : \mathcal{S}(\mathcal{L}) \to \mathbb{R}$, one for each player $i \in N$. $\mathcal{L}$ is called the **underlying language**; we omit it as a prefix when it is safe to do so.

If we extend Γ to an $\mathcal{L}(\Phi_\Gamma)$-game, the players' utility functions are essentially defined on Σ, so an $\mathcal{L}(\Phi_\Gamma)$-game is really just a classical 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, a structured underlying language provides tools that can be useful for studying the corresponding class of language-based games; in particular, it makes it easier to analyze the much broader class of psychological games.

A psychological game is just like a classical 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 [Fagin *et al.*, 1995]. But for many applications of interest, understanding the (completely standard) 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. We conclude this extended abstract with two more examples.

Example 2.1: *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, self-serving behaviour.

The standard prisoner's dilemma is summarized in Table 2:

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

Table 2: 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

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

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 she'll defect, then since her utility in this case would be 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 the full paper that this game has no Nash equilibrium. On the other hand, players can certainly best-respond to their beliefs, and the corresponding iterative notion of *rationalizability* finds purchase here. In the full paper we import this solution concept into our framework and show that every strategy for the indignant altruist is rationalizable. ∎

Example 2.2: *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.

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, say, drive an extra 5 kilometers to save \$50 on a \$300 purchase are often not willing to drive that same extra distance to save the same amount of money on a \$20,000 purchase.

The idea is to assume a certain kind of coarseness, specifically, that the language over which Alice forms preferences does not describe prices with infinite precision. For example, we might assume that the language includes as primitive propositions terms of the form p_Q, where Q ranges over a given partition of the real line, say of the form

$$\cdots \cup [280, 290) \cup [290, 300) \cup [300, 310) \cup \cdots,$$

at least around the \$300 mark. Any utility function defined over such a language cannot distinguish prices that fall into the same partition. Thus, in the example above, Alice would consider the prices \$300 and \$301 to be the same as far as her preferences are concerned. If, moreover, we assume that the partition that determines the underlying language is not only coarse, but is coarser for higher prices, then Alice may very well prefer a price of \$300 to a price of \$350, but not prefer a price of \$20,000 to a price of \$20,050. This has a certain intuitive appeal: the higher numbers get the less precise your langauge is in describing them. Indeed, psychological experiments have demonstrated that Weber's law[1], traditionally applied to physical stimuli, finds purchase in the realm of numerical perception: larger numbers are subjectively harder to discriminate from one another [Moyer and Landauer, 1967; Restle, 1978]. Our choice of underlying language represents this phenomenon simply, while exhibiting its explanatory power. ∎

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 grants FA9550-08-1-0266 and 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.

[1]Weber's law asserts that the minimum difference between two stimuli necessary for a subject to discriminate between them increases as the magnitude of the stimuli increases.

References

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

[Bjorndahl *et al.*, 2013] A. Bjorndahl, J. Y. Halpern, and R. Pass. Language-based games. In *Proceedings of the Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge*, pages 39–48, 2013.

[Fagin *et al.*, 1995] 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.

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

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

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

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

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