Bayesian Games with Intentions

Adam Bjorndahl
Carnegie Mellon University
Philosophy
Pittsburgh, PA 15213
abjorn@andrew.cmu.edu

Joseph Y. Halpern
Cornell University
Computer Science
Ithaca, NY 14853
halpern@cs.cornell.edu

Rafael Pass
Cornell University
Computer Science
Ithaca, NY 14853
rafael@cs.cornell.edu

We show that standard Bayesian games cannot represent the full spectrum of belief-dependent preferences. However, by introducing a fundamental distinction between *intended* and *actual* strategies, we remove this limitation. We define *Bayesian games with intentions*, generalizing both Bayesian games and psychological games [5], and prove that Nash equilibria in psychological games correspond to a special class of equilibria as defined in our setting.

1 Introduction

Type spaces were introduced by John Harsanyi [6] as a formal mechanism for modeling games of incomplete information where there is uncertainty about players' payoff functions. Broadly speaking, types are taken to encode payoff-relevant information, a typical example being how each participant values the items in an auction. An important feature of this formalism is that types also encode beliefs about types. Thus, a type encodes not only a player's beliefs about other players' payoff functions, but a whole *belief hierarchy*: a player's beliefs about other players' beliefs, their beliefs about other players' beliefs about other players' beliefs, and so on.

This latter point has been enthusiastically embraced by the epistemic game theory community, where type spaces have been co-opted for the analysis of games of *complete* information. In this context, types encode beliefs about the *strategies* used by other players in the game as well as their types. So again, types encode a belief hierarchy, but one that describes players' beliefs about other players' beliefs... about the other players' types and strategies. In this framework, one can determine whether a player is rational given her type and strategy; that is, whether her strategy is such that she is making a best response to her beliefs, as encoded by her type. Thus, rationality, common belief of rationality, and so on can be defined as events in the space of (profiles of) strategy-type pairs. This opens the way to epistemic analyses of solution concepts, among other applications [3]. In this setting, types do not encode payoff-relevant information; rather, they are simply a tool for describing belief hierarchies about other players' (types and) strategies.

By contrast, in a *Bayesian game*, types are payoff-relevant objects in that utility depends on them, though the payoff-relevant information they are taken to encode often includes such things as characteristics of the players (strength, work ethic, etc.), or more generally any relevant facts that may not be common knowledge. There is typically assumed to be a prior probability on types (indeed, often a common prior), so a type can still be viewed as encoding beliefs on other types in this setting (a type t encodes the probability obtained by conditioning the prior on t), and thus a belief hierarchy. However, the only aspect of this belief hierarchy that is typically used in Bayesian games is the first-order belief about other players' types (but not beliefs about beliefs, and so on), which is needed to defined a player's expected

Submitted to:
TARK 2015

utility. Nonetheless, it is possible to leverage the fact that types encode beliefs to define Bayesian games in which players' preferences depend to some extent on the *beliefs* of their opponents (see Example 2.2). This observation is the point of departure for the present work.

The notion that a player's preferences might depend on the beliefs of her opponents (or on her own beliefs) is not new. *Psychological games* [1, 5] model phenomena like anger, surprise, and guilt by incorporating belief hierarchies directly into the domain of the utility functions. *Language-based games* [2] model similar belief-dependent preferences by defining utility over descriptions in a given language (in particular, a language that can express the players' beliefs). Types play no explicit role in either of these frameworks; on the other hand, the discussion above suggests that they may be naturally employed to accomplish many of the same modeling goals. Since Bayesian games and, more generally, type spaces have become cornerstones of modern game theory, if the modeling and analysis of psychological games could be carried out in this familiar framework, it would unify these paradigms and thereby amplify both the insights and the accessibility of the latter. In this paper, we provide an extension of Bayesian games that allows us to do just this.

There is an obvious obstruction to capturing general belief-dependent preferences using types in the standard way: types in Bayesian games encode beliefs about types, not about strategies. This severely limits the extent to which preferences over types can capture feelings like surprise or guilt, which are typically expressed by reference to beliefs about strategies (e.g., my opponent is surprised if I do not play the strategy that she was expecting me to play). It may seem that there is a simple solution to this problem: allow types to encode beliefs about strategies. But doing this leads to difficulties in the definition of *Bayesian Nash equilibrium*, the standard solution concept in Bayesian games; this notion depends on being able to freely associate strategies with types. In Section 2, we give the relevant definitions and make this issue precise.

In Section 3, we develop a modification of the standard Bayesian setup where each player is associated with *two* strategies: an *intended* strategy that is determined by her type (and thus can be the object of beliefs), and an *actual* strategy that is independent of her type (as in standard Bayesian games). This gives us what we call *Bayesian games with intentions*. We define a solution concept for such games where we require that, in equilibrium, the actual and intended strategies are equal. As we show, under this requirement, equilibria do not always exist.

In Section 4, we show that psychological games can be embedded in our framework. Moreover, we show that the notion of Nash equilibrium for psychological games defined by Geanakoplos, Pearce, and Stachetti [5] (hereafter GPS) corresponds to a special case of our own notion of equilibrium. Thus, we realize all the advantages of psychological games in an arguably simpler, better understood setting. We do not require complicated beliefs hierarchies; these are implicitly encoded by types.

The advantages of distinguishing actual from intended strategies go well beyond psychological games. As we show in the full paper, intended strategies can be fruitfully interpreted as *reference points* in the style of prospect theory [7]. One of the central insights of prospect theory is that the subjective value of an outcome can depend, at least in part, on how that outcome compares to some "reference level"; for example, whether it is viewed as a relative gain or loss. The intended/actual distinction naturally implements the needed comparison between "real" and "reference" outcomes. Using this insight, we show that *reference-dependent preferences*, as defined by Kőszegi and Rabin [8], can be captured using Bayesian games with intentions.

2 Bayesian games

2.1 Definition

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.

A *type* of player i is often construed as encoding precisely such characteristics. More generally, however, types can be viewed as encoding any kind of information about the world that might be payoff-relevant. For example, the resolution of a battle between two armies may depend not only on what maneuvers they each perform, but also on how large or well-trained they were to begin with, or the kind of terrain they engage on. Decision-making in such an environment therefore requires a representation of the players' uncertainty regarding these variables.

We now give a definition of Bayesian games that is somewhat more general than the standard definition. This will make it easier for us to develop the extension to Bayesian games with intentions. We explain the differences after we give the definition.

Fix a set of *players*, $N = \{1, \ldots, n\}$. A **Bayesian game (over** N**)** is a tuple $\mathscr{B} = (\Omega, (\Sigma_i, T_i, \tau_i, p_i, u_i)_{i \in N})$ where

- Ω is the measurable space of *states of nature*;
- Σ_i is the set of *strategies available to player i*;
- T_i is the set of *types of player i*;
- $\tau_i : \Omega \to T_i$ is *player i's signal function*;
- $p_i : T_i \to \Delta(\Omega)$ associates with each type t_i of player i a probability measure $p_i(t_i)$ on Ω with $p_i(t_i)(\tau_i^{-1}(t_i)) = 1$, representing *type t_i of player i's beliefs* about the state of nature;[1]
- $u_i : \Sigma \times \Omega \to \mathbb{R}$ is *player i's utility function*.[2]

As we said above, this definition of a Bayesian game is more general than what is presented in much (though not all) of the literature. There are two main differences. First, we take utility to be defined over strategies and *states of nature*, rather than over strategies and types (cf. [9] for a similar definition). This captures the intuition that what is really payoff-relevant is *the way the world is*, and types simply capture the players' imperfect knowledge of this. Since the type signal function profile $(\tau_1, \ldots, \tau_n)$ associates with each world a type profile, utilities can depend on players' types. Of course, we can always restrict attention to the special case where $\Omega = T$ and where $\tau_i : T \to T_i$ is the ith projection function; this is called the *reduced form*, and it accords with a common conception of types as encoding all payoff-relevant information aside from strategy choices (cf. [4]).

The second respect in which this definition is more general than is standard is in the association of an *arbitrary* probability measure $p_i(t_i)$ to each type t_i. It is typically assumed instead that for each player i there is some fixed probability measure $\pi_i \in \Delta(\Omega)$ representing her "prior beliefs" about the state of nature, and $p_i(t_i)$ is obtained by conditioning these prior beliefs on the "private information" t_i (or, more

[1] As usual, we denote by $\Delta(X)$ the set of probability measures on the measurable space X. To streamline the presentation, we suppress measurability assumptions here and elsewhere in the paper.

[2] Given a collection $(X_i)_{i \in N}$ indexed by N, we adopt the usual convention of denoting by X the product $\prod_{i \in N} X_i$ and by X_{-i} the product $\prod_{j \neq i} X_j$.

precisely, on the event $\tau_i^{-1}(t_i)$).[3] 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. We adopt the more general setup because it accords with a standard presentation of type spaces as employed for the epistemic analysis of games of complete information [3], thus making it easier for us to relate our approach to epistemic game theory.

The requirement that $p_i(t_i)(\tau_i^{-1}(t_i)) = 1$ amounts to assuming that each player is sure of her own type (and hence, her beliefs); that is, in each state $\omega \in \Omega$, each player i knows that the true state is among those where she is of type $t_i = \tau_i(\omega)$, which is exactly the set $\tau_i^{-1}(t_i)$.

2.2 Examples

It will be helpful to briefly consider two simple examples of Bayesian games, one standard and one a bit less so.

Example 2.1: First consider a simplified auction scenario where each participant $i \in N$ must submit a bid $\sigma_i \in \Sigma_i = \mathbb{R}^+$ for a given item. Types here are conceptualized as encoding valuations of the item up for auction: for each $t_i \in T_i$, let $v(t_i) \in \mathbb{R}^+$ represent how much player i thinks the item is worth, and define player i's utility $u_i : \Sigma \times T$ by

$$u_i(\sigma,t) = \begin{cases} v(t_i) - \sigma_i & \text{if } \sigma_i = \max_{j \in N} \sigma_j \\ 0 & \text{otherwise.} \end{cases}$$

Thus, a player's payoff is 0 if she does not submit the highest bid, and otherwise is equal to her valuation of the item less her bid (for simplicity, this model assumes that in the event of a tie, every top-bidding player gets the item). Note that the state space here is implicitly taken to be identical to the space T of type profiles, that is, the game is presented in reduced form. A type t_i therefore tells us not only how valuable player i thinks the item is ($v(t_i)$), but also what beliefs $p_i(t_i) \in \Delta(T)$ player i has about how the *other* players value the item (and what beliefs they have about *their* opponents, and so on). The condition that $p_i(t_i)(\tau^{-1}(t_i)) = 1$ then simply amounts to the assumption that each player is sure of her own valuation (as well as her beliefs about other players' types). ∎

Example 2.2: Next we consider an example where the Bayesian framework is leveraged to model a player whose preferences depend on the beliefs of her opponent. Consider a game where the players are students in a class, with player 1 having just been called upon by the instructor to answer a yes/no question. Assume for simplicity that $N = \{1,2\}$, $\Sigma_1 = \{\mathsf{yes}, \mathsf{no}, \mathsf{pass}\}$, and $\Sigma_2 = \{*\}$ (where $*$ denotes a vacuous move, so only player 1 has a real decision to make). Let $\Omega = \{w_y, w_n, v_y, v_n\}$, where, intuitively, states with the subscript y are states where "yes" is the correct answer, while states with the subscript n are states where "no" is the correct answer. Let $T_1 = \{t_1, t_1'\}$, $T_2 = \{t_2, t_2', t_2''\}$, and define the signal functions by

$$\tau_1(w_y) = \tau_1(w_n) = t_1,\ \tau_1(v_y) = \tau_1(v_n) = t_1', \text{ and}$$
$$\tau_2(w_y) = \tau_2(w_n) = t_2 \text{ and } \tau_2(v_y) = t_2' \text{ and } \tau_2(v_n) = t_2''.$$

Finally, assume that all of the subjective probability measures arise by conditioning a common prior $\pi \in \Delta(\Omega)$ on the type of the player in question; assume further that π is the uniform distribution. It follows that in each state, player 1 is unsure of the correct answer. On the other hand, while in states

[3] To ensure this is well-defined, it is also typically assumed that none of player i's types are null with respect to π_i; that is, for all $t_i \in T_i$, $\pi_i(\tau_i^{-1}(t_i)) > 0$.

w_y and w_n, player 2 is also unsure of the correct answer, in states v_y and v_n, player 2 knows the correct answer. Moreover, in states w_y and w_n, player 1 is sure that player 2 does not know the correct answer, whereas in states v_y and v_n, player 1 is sure that player 2 *does* know the correct answer (despite not knowing it himself). We can therefore use this framework to encode the following (quite plausible) preferences for player 1: guessing the answer is preferable to passing provided player 2 does not know the right answer, but passing is better than guessing otherwise. Set

$$u_1(\mathsf{yes}, w_y) = u_1(\mathsf{yes}, v_y) = u_1(\mathsf{no}, w_n) = u_1(\mathsf{no}, v_n) = 5,$$

representing a good payoff for answering correctly; set

$$u_1(\mathsf{pass}, x) = -2 \text{ for all } x \in \Omega,$$

representing a small penalty for passing regardless of what the correct answer is; finally, set

$$u_1(\mathsf{yes}, w_n) = u_1(\mathsf{no}, w_y) = -5 \text{ and}$$
$$u_1(\mathsf{yes}, v_n) = u_1(\mathsf{no}, v_y) = -15,$$

representing a penalty for getting the wrong answer that is substantially worse in states where player 2 knows the correct answer.

It is easy to check that if player 1 considers w_y and w_n to be equally probable, then her expected utility for randomly guessing the answer is 0, which is strictly better than passing (passing, of course, always yields an expected utility of -2). By contrast, if player 1 considers v_y and v_n to be equally probable, then her expected utility for randomly guessing is -5, which is strictly worse than passing. In short, player 1's decision depends on what she believes about the beliefs of player 2. ∎

Example 2.2 captures what might be thought of as *embarrassment aversion*, which is a species of belief-dependent preference: player 1's preferences depend on what player 2 believes. It is worth being explicit about the conditions that make this possible:

C1. States in Ω encode a certain piece of information I (in this case, whether the correct answer to the given question is "yes" or "no").

C2. Types encode beliefs about states.

C3. Utility depends on types.

From C1–C3, we can conclude that preferences can depend on what the players believe about I.

Not all kinds of belief-dependent preferences can be captured in the Bayesian framework. Suppose, for example, that the goal of player 1 is to surprise her opponent by playing an unexpected strategy. More precisely, suppose that $\Sigma_1 = \{\sigma_1, \sigma_1'\}$ and we wish to define u_1 in such a way that player 1 prefers to play σ_1 if and only if player 2 believes he will play σ_1'. In contrast to Example 2.2, this scenario cannot be represented with a Bayesian game for the following simple reason: *states do not encode strategies*. In other words, condition C1 is not satisfied if we take I to be player 1's strategy. Therefore, types cannot encode such beliefs about strategies, so utility cannot be defined in a way that depends on such beliefs.

This suggests an obvious generalization of the Bayesian setting, namely, encoding strategies in states. Indeed, this is the idea we explore in this paper; however, it is not quite as straightforward a maneuver as it might appear, primarily due to its interaction with the mechanics of *Bayesian Nash equilibrium*.

2.3 Equilibrium

Part of the value of Bayesian games lies in the fact that a generalized notion of Nash equilibrium can be defined in this framework, for which the following notion plays a crucial role: a *behaviour rule* for player i is a function $\beta_i : T_i \to \Sigma_i$. In Bayesian games, we talk about behaviour rule profiles being in equilibrium, just as in normal-form games, we talk about strategy profiles being in equilibrium. Intuitively, $\beta_i(t_i)$ represents the strategy that type t_i of player i is playing, so a player's strategy depends on her type.

From a technical standpoint, behaviour rules are important because they allow us to associate a payoff for each player with each *state*, rather than strategy-state pairs. Since types encode beliefs about states, this yields a notion of expected utility for each type. A Bayesian Nash equilibrium is then defined to be a profile of behaviour rules such that each type is maximizing its own expected utility.

More precisely, observe that via the signal functions τ_i, a behaviour rule β_i associates with each state ω the strategy $\beta_i(\tau_i(\omega))$. Thus, a profile β of behaviour rules defines an *induced utility function* $u_i^\beta : \Omega \to \mathbb{R}$ as follows:

$$u_i^\beta(\omega) = u_i((\beta_j(\tau_j(\omega)))_{j\in N}, \omega).$$

The beliefs $p_i(t_i)$ then define the *expected utility* for each type: let $E_{t_i}(\beta)$ denote the expected value of u_i^β with respect to $p_i(t_i)$. Denote by B_i the set of all behaviour rules for player i. A behaviour rule β_i is a **best response to** β_{-i} if, for each $t_i \in T_i$, β_i maximizes E_{t_i}:

$$(\forall \beta_i' \in B_i)(E_{t_i}(\beta_i, \beta_{-i}) \geq E_{t_i}(\beta_i', \beta_{-i})).$$

Finally, a **Bayesian Nash equilibrium** of the Bayesian game $\mathscr{B}$ is a profile of behaviour rules β such that, for each $i \in N$, β_i is a best response to β_{-i}. A (mixed) Bayesian Nash equilibrium is guaranteed to exist when the strategy and types spaces are finite (see [10] for a more general characterization of when an equilibrium exists).

3 Intention

3.1 Definition

Behaviour rules map types to strategies, but the underlying model does not enforce any relationship between types and strategies (or between states and strategies). Thus, behaviour rules do not provide a mechanism satisfying condition C1 with I taken to be a player's strategies, so they do not allow us to express preferences that depend on beliefs about strategies. In order to express such preferences, we must associate strategies with states in the model itself. Note that once we do this, utility functions depend on strategies in two ways. Specifically, since u_i is defined on the cross product $\Sigma \times \Omega$, players' preferences depend on strategies both directly (corresponding to the strategy-profile component of u_i's input) and as encoded in states (the second component of u_i's input). To keep track of this distinction, we call these *actual* and *intended* strategies, respectively.

Formally, a **Bayesian game with instantiated intentions** (BGII) is a tuple $\mathscr{I} = (\Omega, (\Sigma_i, T_i, \tau_i, s_i, p_i, u_i)_{i\in N})$, where $s_i : T_i \to \Sigma_i$ is *player i's intention function* and the remaining components are defined as in a Bayesian game. (The reason for this terminological mouthful will become clear in Section 3.3, where we define *Bayesian games with intentions*.) Each s_i associates 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 "intends" or "is planning" to play (though may ultimately decide not to); alternatively, it might be conceptualized as the "default" strategy for that type; it might even be viewed as the "stereotypical" strategy employed by players of type t_i. The former interpretation may be appropriate in a situation where we want to think of self-control; for example, a player who intends to exercise, but actually does not. The latter interpretation may be appropriate if we think about voting. Wealthy people in Connecticut typically vote Republican, but a particular player i who is wealthy and lives in Connecticut (this information is encoded in her type) votes Democrat.

We associate intended strategies with types rather than directly with states by analogy to behaviour rules, in keeping with the modeling paradigm where the personal characteristics of a player—including her beliefs, decisions, *and intentions*—are entirely captured by her type. Nonetheless, the composition $s_i \circ \tau_i : \Omega \to \Sigma_i$ does associate strategies with states and so satisfies condition C1 (again, with I being a player's strategy); thus, players can have beliefs about strategies. This, in turn, allows us to define utility so as to capture preferences that depend on beliefs about strategies.

3.2 Examples

The presentation of a BGII is made clearer by introducing the following notation for the set of states where player i intends to play σ_i:

$$[\![\sigma_i]\!] = (s_i \circ \tau_i)^{-1}(\sigma_i) = \{\omega \in \Omega : s_i(\tau_i(\omega)) = \sigma_i\}.$$

Example 3.1: Consider a 2-player game in which player 1's goal is to surprise her opponent. We take player 2 to be surprised if his beliefs about what player 1 intends to play are dramatically different from what player 1 actually plays. For definiteness, we take "dramatically different" to mean that his beliefs about player 1's intended strategy ascribe probability 0 to player 1's actual strategy. Thus, we define player 1's utility function as follows:

$$u_1(\sigma, \omega) = \begin{cases} 1 & \text{if } p_2(\tau_2(\omega))([\![\sigma_1]\!]) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

(Recall that $p_2(\tau_2(\omega))$ is a measure on states, which is why we apply it to $\tau_1^{-1}(s_1^{-1}(\sigma_1))$, that is, the set of states ω where player 1's intended strategy, $s_1(\tau_1(\omega))$, is equal to σ_1.) ∎

Example 3.2: Next we consider an example introduced by GPS [5] called *the bravery game*. This is a 2-player scenario in which player 1 has the only real decision to make: he must choose whether to take a *bold* action or a *timid* action, so $\Sigma_1 = \{\mathsf{bold}, \mathsf{timid}\}$ (and $\Sigma_2 = \{*\}$). The crux of the game is the psychological factor, described by GPS as follows: player 1 prefers "to be timid rather than bold, unless he thinks his friends expect him to be bold, in which case he prefers not to disappoint them" [5]. It is also stipulated that player 2 prefers player 1 to be bold, and also prefers to think of him as bold. Define $q : T \to [0,1]$ by

$$q(t) = p(t_2)([\![\mathsf{bold}]\!]),$$

and $\tilde{q} : T \to [0,1]$ by

$$\tilde{q}(t) = E_{t_1}(q),$$

where $E_{t_i}(f)$ denotes the expected value of f with respect to the measure $p(t_i)$. We can then represent the players' preferences in a reduced-form BGII as follows:

$$u_1(\sigma, t) = \begin{cases} 2 - \tilde{q}(t) & \text{if } \sigma_1(t_1) = \mathsf{bold} \\ 3(1 - \tilde{q}(t)) & \text{if } \sigma_1(t_1) = \mathsf{timid}, \end{cases}$$

$$u_2(\sigma,t) = \begin{cases} 2(1+q(t)) & \text{if } \sigma_1(t_1) = \mathsf{bold} \\ 1-q(t) & \text{if } \sigma_1(t_1) = \mathsf{timid}. \end{cases}$$

This representation closely parallels that given in [5], in which q and $\tilde{q}$ are understood not as functions of types, but (implicitly) as functions of belief hierarchies.[4] But this makes no difference to the preferences this game encodes. For example, it is easy to see that player 2 prefers player 1 to be bold, and all the more so when q is high—that is, all the more so when she believes with high probability that he will be bold.[5] Similarly, one can check that player 1 prefers to be timid provided that $\tilde{q}(t) < \frac{1}{2}$; in other words, provided that his expectation of his opponent's degree of belief in him being bold is sufficiently low.

Why not define player 1's preferences directly in terms of the beliefs of his opponent, rather than his expectation of these beliefs? GPS cannot do so because of a technical limitation of the framework as developed in [5]; specifically, that a player's utility can depend only on *her own* beliefs. Battigalli and Dufwenberg [1] correct this deficiency. BGIIs do not encounter such limitations in the first place. In particular, it is easy enough to redefine player 1's utility as follows:

$$u'_1(\sigma,t) = \begin{cases} 2-q(t) & \text{if } \sigma_1(t_1) = \mathsf{bold} \\ 3(1-q(t)) & \text{if } \sigma_1(t_1) = \mathsf{timid}. \end{cases}$$

In this case, we find that player 1 prefers to be timid provided $q(t) < \frac{1}{2}$, or in other words, provided that his opponent's degree of belief in him being bold is sufficiently low. ∎

Observe that in neither of the preceding examples did we provide a concrete BGII, in that we did not explicitly define the type spaces, the intention functions, and so on. Instead, we offered general recipes for implementing certain belief-dependent preferences (e.g., to surprise, to live up to expectations, etc.) in arbitrary BGIIs. Particular choices of type spaces and intention functions do play an important role in equilibrium analyses; however, as illustrated by the preceding two examples, at the modeling stage they need not be provided up front.

3.3 Equilibrium

We now define a notion of equilibrium for this setting. As a first step towards this definition, given a BGII $\mathscr{I}$, we say that a profile of behaviour rules β is an **equilibrium of** $\mathscr{I}$ provided:

(1) β is a Bayesian Nash equilibrium of the underlying Bayesian game: that is, each β_i is a best response to β_{-i} in precisely the sense defined in Section 2.3;

(2) for each player $i \in N$, $\beta_i = s_i$.

This definition, and in particular condition (2), embodies the conception of equilibrium as a steady state of play where each player has correct beliefs about her opponents (and is best responding to those beliefs). In a BGII, beliefs about the strategies of one's opponents are beliefs about intended strategies (although, in equilibrium, a player will also have beliefs about actual strategies). On the other hand,

[4] Additionally, GPS give the value of q, not by the probability that player 2 assigns to player 1 being bold, but by player 2's *expectation* of the probability p with which player 1 decides to be bold. We forgo this subtlety for the time being.

[5] It is not quite clear why GPS define player 2's payoff in the event that player 1 is timid to be $1-q(t)$ rather than $1+q(t)$. This latter value preserves the preferences described while avoiding the implication that, assuming that player 1 will be timid, player 2 also prefers to *believe* that he will be timid—this stands in opposition to the stipulation that player 2 prefers to think of her opponent as bold.

since behavior rules associate strategies with types and players have beliefs over types, behaviour rules also induce beliefs about strategies; in our terminology, these are beliefs about actual strategies. Condition (2) implies that these two beliefs coincide in equilibrium; in equilibrium, each type of each player actually plays the strategy she intended to play (which is exactly the strategy her opponents expected her to play).

Does condition (2) collapse the distinction between intended and actual strategies, thereby returning us to the classical setting? It does not. First, in a standard Bayesian game we could not even write down a model where players' preferences depended on beliefs about strategies. In addition, although we demand that intended and actual strategies coincide in equilibrium, this restriction *does not apply to the evaluation of best responses*. Recall that β_i is a best response to β_{-i} if and only if

$$(\forall \beta_i' \in B_i)(E_{t_i}(\beta_i, \beta_{-i}) \geq E_{t_i}(\beta_i', \beta_{-i})).$$

Crucially, β_i' need not be equal to s_i. In other words, for β_i to count as a best response, it must be at least as good as all other behaviour rules, including those that recommend playing a strategy distinct from that specified by s_i.

Example 3.3: Consider a 2-player reduced-form BGII with $\Sigma_1 = \{a,b\}$, $\Sigma_2 = \{*\}$, $T_1 = \{x,x'\}$, and $T_2 = \{y,y'\}$, and where

$$p_1(x)(\{y\}) = p_1(x')(\{y'\}) = p_2(y)(\{x'\}) = p_2(y')(\{x\}) = 1.$$

Let u_1 be defined as in Example 3.1, encoding player 1's desire to surprise her opponent:

$$u_1(\sigma_1, *, t) = \begin{cases} 1 & \text{if } p_2(t_2)([\![\sigma_1]\!]) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that $s_1(x) = s_1(x') = a$. Then, of course, $p_2(y)([\![a]\!]) = p_2(y')([\![a]\!]) = 1$, and likewise $p_2(y)([\![b]\!]) = p_2(y')([\![b]\!]) = 0$. It follows immediately that the expected utility of playing a for either type of player 1 is equal to 0 (since player 1 is sure that this will not surprise her opponent), whereas the expected utility of playing b for either type of player 1 is equal to 1 (since, in this case, player 1 is sure that this *will* surprise her opponent). In particular, if $\beta_1 = s_1$, then β_1 is not a best response. Thus, this particular BGII admits no equilibrium.

Now suppose that $s_1(x) = a$ and $s_1(x') = b$. This is, of course, a different BGII from the one considered in the previous paragraph, but it differs only in the specification of player 1's intentions. Moreover, in this BGII it is not hard to check that $\beta_1 = s_1$ is a best response and therefore constitutes an equilibrium: type x is sure that player 2 is of type y; therefore, type x is sure that player 2 is sure that player 1 is of type x', and so is playing b; thus, a is a best response for x, since x is sure that it will surprise her opponent; a similar argument shows that b is a best response for x'. ∎

Example 3.3 demonstrates that the notion of best response in a BGII—and therefore the notion of equilibrium—can be sensitive to states of play where players are *not* playing their intended strategies. But it also illustrates the pivotal role of the intention functions s_i in determining the existence of an equilibrium. Indeed, condition (2) implies that if a given BGII $\mathscr{I}$ has an equilibrium at all, it is unique and equal to s. This suggests that BGIIs are not at the right "resolution" for equilibrium analysis, since they come already equipped with a unique candidate for equilibrium. Thus, rather than restricting attention to a single BGII, where the intention function is specified and hard-coded into the model, we consider a more general model, where the intention function is not specified, but still affects the utility. This is parallel to the role of strategies in standard games, which are not hard-coded into the model, but of course the utility function still depends on them. Essentially, we are moving the intention function from the model to the utility function. As we shall see, our earlier examples of BGIIs can be easily interpreted as models in this more general sense.

In order to make this precise, we must first formally define utility functions that take as arguments intention functions. More precisely, taking $\Sigma^T = \Sigma_1^{T_1} \times \cdots \times \Sigma_n^{T_n}$ (so that Σ^T is the set of intention function profiles), an *explicit utility function* is a map $\tilde{u}_i : \Sigma \times \Omega \times \Sigma^T \to \mathbb{R}$; these are just like the utility functions in a BGII except they explicitly take as input the associations between types and strategies provided by intention functions. A **Bayesian game with intentions** (BGI) is a tuple $\tilde{\mathscr{I}} = (\Omega, (\Sigma_i, T_i, \tau_i, p_i, \tilde{u}_i)_{i \in N})$, where the components are defined just as they are in a Bayesian game, except that the functions $\tilde{u}_i$ are explicit utility functions. We emphasize that a BGI, unlike a BGII, does not include players' intention functions among its components; instead, these functions show up as arguments in the (explicit) utility functions.

It is easy to see that all the examples of BGIIs that we have considered so far can be naturally converted to BGIs. For example, the utility function $u_1(\sigma,t)$ in Example 3.1 becomes $\tilde{u}_i(\sigma,t,s)$. The definition of $\tilde{u}_i(\sigma,t,s)$ looks identical to that of $u_1(\sigma,t)$; the additional argument s is needed to define $[\![\sigma_1]\!]$.

A BGI induces a natural map from intention functions to BGIIs: given $\tilde{\mathscr{I}} = (\Omega, (\Sigma_i, T_i, \tau_i, p_i, \tilde{u}_i)_{i \in N})$ and functions $s_i : T_i \to \Sigma_i$, let

$$\tilde{\mathscr{I}}(s_1, \ldots, s_n) = (\Omega, (\Sigma_i, T_i, \tau_i, s_i, p_i, u_i)_{i \in N}),$$

where $u_i : \Sigma \times \Omega \to \mathbb{R}$ is defined by

$$u_i(\sigma, \omega) = \tilde{u}_i(\sigma, \omega, s_1, \ldots, s_n).$$

Clearly $\tilde{\mathscr{I}}(s_1, \ldots, s_n)$ is a BGII; we call it an **instantiation of** $\tilde{\mathscr{I}}$. We then define an **equilibrium of** $\tilde{\mathscr{I}}$ to be a profile of behaviour rules β that is an equilibrium of the corresponding instantiation $\tilde{\mathscr{I}}(\beta)$. Here we make implicit use of the fact that both behaviour rules and intention functions are functions from types to strategies. Indeed, the profile β plays two roles: first, it is used to determine the intentions of the players; then, in the context of the instantiated BGI with these fixed intentions, we evaluate whether each $\beta_i(t_i)$ is a best response, just as in the definition of equilibrium for a standard Bayesian game.

Is this a reasonable notion of equilibrium? As we observed above, in a BGII, the only possible equilibrium is "built in" to the model in the form of the intention functions. In particular, the only possible equilibrium for the instantiation $\tilde{\mathscr{I}}(\beta)$ is β itself. Of course, β is not necessarily an equilibrium of this game; however, by quantifying over β and considering the corresponding class of BGIIs (i.e., those obtained as instantiations of $\tilde{\mathscr{I}}$), we are essentially asking the question: "Is there a profile of intentions such that, assuming those intentions are common knowledge, no player prefers to deviate from their intention?" If so, that profile constitutes an equilibrium. This is a natural solution concept; in fact, as we show in Section 4, the notion of equilibrium proposed by GPS for psychological games is a special case of our definition.

Example 3.4: In light of these definitions, Example 3.3 can be viewed as first defining a BGI $\tilde{\mathscr{I}}$, and then considering two particular instantiations of it. The equilibrium observations made then amount to the following: the behaviour rule $\beta_1 \equiv a$ (i.e., the constant function a) is not an equilibrium of $\tilde{\mathscr{I}}$, but the behaviour rule β_1' that sends x to a and x' to b is. (As there is only ever one option for player 2's behaviour rule, namely $\beta_2 \equiv *$, we can safely neglect it.) ∎

Example 3.5: Consider again the bravery game of Example 3.2. Under any particular specification of state space and type spaces, this becomes a BGI $\tilde{\mathscr{I}}$. It is not difficult to see that each of the behaviour rules $\beta_1 \equiv \mathsf{timid}$ and $\beta_1' \equiv \mathsf{bold}$ is an equilibrium of $\tilde{\mathscr{I}}$. ▮

3.4 Existence

Are equilibria of BGIs guaranteed to exist? Not necessarily. At least one obstacle to existence lies in the specification of the underlying type space and the corresponding probability measures: as the following example shows, certain kinds of belief that are necessary for best-responses may be implicitly ruled out.

Example 3.6: Consider a 2-player reduced-form BGI $\tilde{\mathscr{I}}$ where $\Sigma_1 = \{a,b\}$, $\Sigma_2 = \{*\}$, $T_1 = \{x,x'\}$, and $T_2 = \{y,y'\}$, and where

$$p_1(x)(\{y\}) = p_1(x')(\{y'\}) = p_2(y)(\{x\}) = p_2(y')(\{x'\}) = 1.$$

Once again we consider a model where player 1 wishes to surprise her opponent, and so define u_1 as in Example 3.3:

$$u_1(\sigma_1,*,t) = \begin{cases} 1 & \text{if } p_2(t_2)([\![\sigma_1]\!]) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that player 1 is certain that player 2 knows her type. It follows that no matter what her intentions are, player 2 knows them, and so (by definition of u_1), player 1 can always do better by deviating. In other words, no behaviour rule β_1 is an equilibrium of $\tilde{\mathscr{I}}(\beta_1)$ (since it is not a best response). It follows immediately that $\tilde{\mathscr{I}}$ admits no equilibria. ▮

This obstacle persists even if we extend our attention to mixed strategies. More precisely, consider the class of BGIIs where, for each player i, $\Sigma_i = \Delta(A_i)$ for some finite set A_i (the set of player i's *pure strategies*), and $u_i : \Sigma \times \Omega \to \mathbb{R}$ satisfies

$$u_i(\sigma_i, \sigma_{-i}, \omega) = \sum_{a_i \in A_i} \sigma_i(a_i) u_i(a_i, \sigma_{-i}, \omega).$$

In other words, player i's utility for playing σ_i is just the expected value of her utility for playing her various pure strategies with the probabilities given by σ_i. As is standard, we call elements of Σ_i *mixed strategies*, and the corresponding BGIIs *mixed-strategy BGIIs*. We can similarly define *mixed-strategy BGIs*. Note that in this context, since the intention functions s_i map into Σ_i, intended strategies are also mixed.

The next example shows that, in contrast to the classical setting, there are mixed-strategy BGIs with finite type spaces that admit no equilibria.

Example 3.7: Consider a 2-player reduced-form BGI where $\Sigma_1 = \Delta(\{a,b\})$, $\Sigma_2 = \{*\}$, $T_1 = \{x,x'\}$, and $T_2 = \{y,y'\}$, and where

$$p_1(x)(\{y\}) = p_1(x')(\{y'\}) = p_2(y)(\{x\}) = p_2(y')(\{x'\}) = 1.$$

Set

$$u_1(a,*,t) = \begin{cases} 1 & \text{if } p_2(t_2)([\![a]\!]) < 1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$u_1(b,*,t) = \begin{cases} 1 & \text{if } p_2(t_2)([\![a]\!]) = 1 \\ 0 & \text{otherwise,} \end{cases}$$

and extend to all $\sigma_1 \in \Delta(\{a,b\})$ by taking expectation:

$$u_1(\sigma_1,*,t) = \sigma_1(a)u_1(a,*,t) + \sigma_1(b)u_1(b,*,t).$$

Note that, following standard conventions, here we identify the pure strategy a with the degenerate mixed strategy that places probability 1 on a; likewise for b. Thus, for example, the condition $p_2(t_2)([\![a]\!]) < 1$ amounts to the following: "type t_2 is not absolutely certain that player 1 intends to play the pure strategy a", or equivalently, "type t_2 considers it possible that player 1 intends to play a mixed strategy that places positive weight on b". The preferences defined by u_1 can be roughly summarized as follows: "player 1 prefers to play a in the event that player 2 thinks she might place positive weight on b, and prefers to play b if player 2 is sure that she'll play a for sure".

This game admits no equilibria. To see this, suppose that β_1 were an equilibrium: that is, set player 1's intention function equal to β_1, and suppose that β_1 is an equilibrium of the resulting BGII.[6] First consider the case where $\beta_1(x) \in \Sigma_1$ satisfies $\beta_1(x)(b) > 0$. Then it follows that $p_2(y)([\![a]\!]) = 0$ (i.e., type y is certain that player 1 is not playing the pure strategy a), and so, since type x is certain that player 2 is of type y, it follows by definition of u_1 that type x's best response is to play the pure strategy a. In particular, $\beta_1(x)$ is not a best response, so β_1 cannot constitute an equilibrium. Now consider the case where $\beta_1(x)(b) = 0$; in other words, $\beta_1(x)$ is the pure strategy a. Then we have $p_2(y)([\![a]\!]) = 1$, from which it follows that type x's best response is to play the pure strategy b. Thus, once again, β_1 cannot constitute an equilibrium. ∎

4 Psychological games

Psychological games can be captured in our framework. A psychological game $\mathscr{P}$ consists of a finite set of players N, together with mixed strategies Σ_i and utility functions $v_i : \bar{B}_i \times \Sigma \to \mathbb{R}$ for each player i, where $\bar{B}_i$ denotes the set of "collectively coherent" belief hierarchies for player i. Somewhat more precisely, an element $b_i \in \bar{B}_i$ is an infinite sequence of probability measures $(b_i^1, b_i^2, \dots)$ where $b_i^1 \in \Delta(\Sigma_{-i})$ is player i's *first-order beliefs*, b_i^2 is player i's *second-order beliefs* (i.e., roughly speaking, her beliefs about the beliefs of her opponents), and so on, such that the beliefs in this sequence satisfy certain technical conditions (roughly speaking, lower-order beliefs must agree with the appropriate marginals of higher-order beliefs, and this agreement must be common knowledge); see the full paper for the complete definition.

Given a mixed-strategy BGII $\mathscr{I}$ and a type $t_i \in T_i$, we can define the first-order beliefs associated with t_i by

$$\varphi_i^1(t_i) = (s_{-i})_*(p_i(t_i));$$

that is, the pushforward of $p_i(t_i)$ from Ω to Σ_{-i} by s_{-i}. Note that, in our terminology, these are beliefs about *intended* strategies. The kth-order beliefs associated with t_i, denoted $\varphi_i^k(t_i)$, can be defined inductively in a similar fashion; it is then straightforward to show that the sequence

$$\varphi_i(t_i) = (\varphi_i^1(t_i), \varphi_i^2(t_i), \dots)$$

[6] As before, we ignore player 2's behaviour since he has no choices to make.

is collectively coherent, and thus $\varphi_i : T_i \to \bar{B}_i$ (see the full paper).

This correspondence provides a natural notion of equivalence between psychological games and BGIIs with respect to the psychological preferences expressed in the former, namely,

$$\forall i \in N\, \forall \sigma \in \Sigma \forall \omega \in \Omega(u_i(\sigma,\omega) = v_i(\varphi_i(\tau_i(\omega)),\sigma)).$$

When a BGII $\mathscr{I}$ satisfies this condition with respect to a psychological game $\mathscr{P}$, we say that $\mathscr{I}$ and $\mathscr{P}$ are **preference-equivalent**.

The notion of preference-equivalence lifts naturally to BGIs. Observe that the functions φ_i^k depend on the profile of intention functions s; being explicit about this dependence, we write $\varphi_i^k(t_i;s)$ rather than $\varphi_i^k(t_i)$; we then say that $\tilde{\mathscr{I}}$ and $\mathscr{P}$ are preference-equivalent provided that

$$\forall i \in N \forall \sigma \in \Sigma \forall \omega \in \Omega \forall s \in \Sigma^T (\tilde{u}_i(\sigma,\omega,s) = v_i(\varphi_i(\tau_i(\omega);s),\sigma)).$$

It is easy to see that, given a psychological game $\mathscr{P}$, we can obtain a preference-equivalent BGI $\tilde{\mathscr{I}}$ simply by taking the above condition as the *definition* of the utility functions $\tilde{u}_i$. Thus, we have the following:

Proposition 4.1: *For every psychological game there exists a preference-equivalent BGI.*

Note that even very simple BGIs (i.e., those with very small type/state spaces) can be preference-equivalent to psychological games; indeed, it is sufficient for the utility functions $\tilde{u}_i$ to be of the form

$$\tilde{u}_i(\sigma,\omega,s) = f(\varphi_i(\tau_i(\omega);s),\sigma),$$

so that utility depends on states only to the extent that states encode belief hierarchies. In particular, although the utility functions in a psychological game have uncountable domains (since they apply to all possible belief hierarchies), a BGI $\tilde{\mathscr{I}}$ can be preference-equivalent to a psychological game $\mathscr{P}$ even if $\tilde{\mathscr{I}}$ has only finitely many states, since all that matters is that the utility functions of $\tilde{\mathscr{I}}$ agree with the utilitiy functions of $\mathscr{P}$ on the belief hierarchies encoded by the states of $\tilde{\mathscr{I}}$. Given a psychological game, we can construct a preference-equivalent BGI with type spaces rich enough that each φ_i is surjective: in other words, every belief hierarchy is realized by some type. However, in order to capture *equilibrium* behaviour, such richness turns out to be superfluous. We now show how the notion of equilibrium defined by GPS for psychological games can be recovered as equilibria in our setting.

Given $\sigma \in \Sigma$, let $\chi_i(\sigma) \in \bar{B}_i$ denote the unique belief hierarchy for player i corresponding to common belief in σ. A *psychological Nash equilibrium* of $\mathscr{P}$ is a strategy profile σ such that, for each player i, σ_i maximizes the function

$$\sigma_i' \mapsto v_i(\chi_i(\sigma),\sigma_i',\sigma_{-i}).$$

In particular, to check whether σ constitutes a psychological Nash equilibrium, the only relevant belief hierarchies are those corresponding to common belief of σ. This, in essense, is the reason we do not need rich type spaces in BGIs to detect such equilibria.

Theorem 4.2: *If $\mathscr{P}$ and $\tilde{\mathscr{I}}$ are preference-equivalent, then σ is a psychological Nash equilibrium of $\mathscr{P}$ if and only if the profile of (constant) behaviour rules β for which $\beta_i \equiv \sigma_i$ is an equilibrium of $\tilde{\mathscr{I}}$.*

Proof: When β is the profile of behaviour rules described in this theorem, the corresponding instantiation $\tilde{\mathscr{I}}(\beta)$ has the property that, for each type t_i, $\varphi_i(t_i) = \chi_i(\sigma)$. The rest of the proof is essentially just unwinding definitions; see the full paper for details. ∎

Theorem 4.2 shows that equilibrium analysis in psychological games does not depend on the full space of belief hierarchies; it can be captured by particularly simple BGIs. It also establishes an equivalence between psychological Nash equilibria and a certain restricted class of equilibria in BGIs; namely, those consisting of constant behaviour rules. This restriction is not surprising in light of the fact that psychological games do not model strategies as functions of types, while BGIs do. Thus, BGIs are not merely recapitulations of the GPS framework: they are a common generalization of psychological games and Bayesian games.

5 Conclusion

We have introduced BGIs, Bayesian games with intentions, which generalize Bayesian games and psychological games in a natural way. We believe that BGIs will prove much easier to deal with than psychological games, while allowing greater flexibility.

When do equilibria exist? While Theorem 4.2 provides sufficient conditions for the existence of equilibria in BGIs, they are certainly not necessary conditions. We can show, for example, that there are BGIs that admit only equilibria in which no behaviour rule is constant. Formulating more general conditions sufficient for existence is the subject of ongoing work.

Perhaps the most exciting prospect for future research lies in leveraging the distinction between actual and intended strategies. As we show in the full paper, this distinction can be used to implement Köszegi and Rabin's [8] model of reference-dependent preferences; we believe that it will have other uses as well, and perhaps lead to new insights into solution concepts.

Acknowledgements

We are indebted to Aviad Heifetz for asking the question that first prompted us to explore the notion of extending Bayesian games to capture belief-dependent preferences. Halpern is supported in part by NSF grants IIS-0911036 and CCF-1214844, by ARO grant W911NF-14-1-0017, and by the Multidisciplinary University Research Initiative (MURI) program administered by the AFOSR under grant FA9550-12-1-0040. Pass is supported in part by a Microsoft Research Faculty Fellowship, NSF CAREER Award CCF-0746990, AFOSR Award FA9550-08-1-0197, and BSF Grant 2006317.

References

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

[2] A. Bjorndahl, J. Y. Halpern & R. Pass (2013): *Language-based games*. In: Theoretical Aspects of Rationality and Knowledge: Proc. Fourteenth Conference (TARK 2013), pp. 39–48.

[3] E. Dekel & M. Siniscalchi (2015): *Epistemic Game Theory*. In H. P. Young & S. Zamir, editors: Handbook of Game Theory with Economic Applications, Volume 4, North-Holland, pp. 619–702.

[4] D. Fudenberg & J. Tirole (1991): *Game Theory*. MIT Press, Cambridge, Mass.

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

[6] J. Harsanyi (1968): *Games with incomplete information played by 'Bayesian' players, Parts I–III*. Management Science 14, pp. 159–182, 320–334, 486–502.

[7] D. Kahneman & A. Tversky (1979): *Prospect theory, an analysis of decision under risk*. Econometrica 47(2), pp. 263–292.

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

[9] M. J. Osborne & A. Rubinstein (1994): *A Course in Game Theory*. MIT Press, Cambridge, Mass.

[10] G. Tian (2009): *The existence of equilibria in games with arbitrary strategy spaces and payoffs: a full characterization*. Available at www.dklevine.com/archive/refs4814577000000000160.pdf.