Bayesian Games with Intentions[1]

Adam Bjorndahl

Carnegie Mellon University, Pittsburgh, PA, USA

Joseph Y. Halpern

Cornell University, Ithaca, NY, USA

Rafael Pass

Cornell University, Ithaca, NY, USA

Abstract

We define *Bayesian games with intentions* by introducing a distinction between "intended" and "actual" actions, generalizing both Bayesian games and (static) *psychological games* [1]. We propose a new solution concept for this framework and prove that Nash equilibria in static psychological games correspond to a special class of equilibria as defined in our setting. We also show how the actual/intended divide can be used to implement the distinction between "real" outcomes and "reference" outcomes so crucial to *prospect theory*, and how some of the core insights of prospect theory can thereby be captured using Bayesian games with intentions.

1. Introduction

Type spaces were introduced by John Harsanyi [2] 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 (see Example 1). An important feature of this formalism is that types also encode beliefs about types. Thus, a type captures 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.

In a *Bayesian game*, utility functions depend on types as well as actions. In this context, types are often used to encode payoff-relevant features of the players themselves, such as their strength or work ethic; more generally, any relevant

[1] A preliminary version of this paper appears in the *Proceedings of the Fifteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, 2015.

facts that are not common knowledge are fair game. Each player is assumed to have a prior probability on types (indeed, a common prior is often assumed), so types can still be viewed as encoding beliefs about other types in this setting (a type t_i for player i encodes the probability obtained by conditioning player i's prior on t_i), and thus a belief hierarchy. However, the only parts of these belief hierarchies that are typically used in a Bayesian game are the first-order beliefs they encode, namely beliefs about other players' types (but not beliefs about beliefs, etc.). This is because first-order beliefs are needed in order to define a player's expected utility (since utility depends on types). Nonetheless, it is possible to leverage the fact that types encode beliefs to define Bayesian games in which players' preferences depend, in a limited way, on the *beliefs* of their opponents (see Example 2). This observation, and the precise nature of the limitation just mentioned, form 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 by no means new: *psychological games*, beginning with the groundbreaking work of Geanakopolos, Pearce, and Stacchetti [1] (hereafter GPS) and substantially extended by Battigalli and Dufwenberg [3] (hereafter BD) to the dynamic setting, have been applied to model phenomena like anger, surprise, guilt, anxiety, and loss aversion by incorporating belief hierarchies directly into the domain of the utility functions. Types play no explicit role in this framework; on the other hand, the discussion above suggests that they may be naturally extended to accomplish some 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 even partially carried out in this familiar framework, it would be a step toward unifying these paradigms, potentially amplifying 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, but not about actions. 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 actions (e.g., my opponent is surprised if I do not choose the action that she was expecting me to choose). It may seem that there is a simple solution to this problem: expand types to encode beliefs about actions. But doing so 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 actions 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* actions: an "intended" action that is determined by her type (and thus can be the object of beliefs), and an "actual" action that is independent of her type (as in standard Bayesian games). This gives us what we call *Bayesian games with intentions* (BGIs). We define a solution concept for such games where we require that, in equilibrium, the actual and intended actions coincide. As we show, under this requirement, equilibria

do not always exist.

It is important to be clear that, despite the name "intentions", the key contribution of this work is not a novel theory of intentions in strategic games. For one thing, such a theory has already been developed by BD and in subsequent work within the framework of dynamic psychological games [4, 5], where intentions are conceived of as depending on beliefs as well as strategic plans. By contrast, we consider "intended actions" to be primitive—existing and defined solely in contrast to "actual actions". Of course, we did not choose the word "intention" randomly—we believe that the conceptual distinction between intended and actual actions represented in BGIs can be fruitfully understood through the commonsense conception of "intention". But, fundamentally, our contribution lies not in giving an account of intention, but rather in showing how carefully distinguishing two sorts of actions in Bayesian games gives rise to a more general modeling paradigm with many applications.

In Section 4, we show that static psychological games, as defined by GPS, can be embedded in our framework. Moreover, we show that the notion of Nash equilibrium for psychological games defined by GPS [1] corresponds to a special case of our notion of equilibrium. Thus, we realize all the advantages of static psychological games in an arguably simpler, better understood setting. We do not require complicated beliefs hierarchies; these are implicitly encoded by types.

In fact, the advantages of distinguishing actual from intended actions go beyond static psychological games. In Section 5, we show that intended actions can be fruitfully interpreted as "reference points", and thereby used to provide a model of asymmetric preferences on gains and losses, in the sense of prospect theory [6]. 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 point; for example, whether it is viewed as a relative gain or loss. The actual/intended distinction naturally implements the distinction between "real" and "reference" outcomes. Kőszegi and Rabin [7] (hereafter KR) have proposed a concrete model of reference-dependent preferences that identifies reference points with players' expectations. BD also discuss this model and how it can be subsumed within dynamic psychological games. As a concrete illustration of our proposal, we show that the insights of the KR framework also arise in the (static!) BGI framework, and moreover that the definition of *personal equilibrium* proposed by KR corresponds in a natural way to our own notion of equilibrium.

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 that 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.

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

- Ω is the measurable space of states of nature;
- A_i is the set of actions available to player i;
- T_i is the set of types of player i;
- $\tau_i : \Omega \to T_i$ is player i's *type-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 satisfying $p_i(t_i)(\tau_i^{-1}(t_i)) = 1$, representing type t_i of player i's beliefs about the state of nature;[2]
- $u_i : A \times \Omega \to \mathbb{R}$ is player i's utility function.[3]

This definition is somewhat different from what is presented in much (though not all) of the literature. There are two main differences. First, we take utility to be defined over actions and states of nature, rather than over actions and types (Osborne and Rubinstein [8] use 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 action choices (cf. [9]).

The second difference is in the association of an *arbitrary* probability measure $p_i(t_i)$ to each type t_i. Typically, for each player i there is given 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

[2] 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.

[3] 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 $\prod_{j \neq i} X_j$.

the "private information" t_i (or, more precisely, on the event $\tau_i^{-1}(t_i)$).[4] 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 flexible notation because it accords with a standard presentation of type spaces as employed for the epistemic analysis of games of complete information [10], 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 1. First consider a simplified auction scenario where each participant $i \in N$ must submit a bid $a_i \in A_i = \mathbb{R}^+$ 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 $u_i : A \times T$ by

$$u_i(a,t) = \begin{cases} v(t_i) - a_i & \text{if } a_i = \max\limits_{j \in N} a_j \\ 0 & \text{otherwise.} \end{cases}$$

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. 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 $N = \{1, 2\}$, $A_1 = \{\mathsf{yes}, \mathsf{no}, \mathsf{pass}\}$, and $A_2 = \{*\}$ (where $*$ denotes a vacuous move, so only player 1 has a real decision

[4] 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 $t_i \in T_i$, $\pi_i(\tau_i^{-1}(t_i)) > 0$.

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 type-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 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 that 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 ascribes probability $1/2$ to each of w_y and w_n, 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 ascribes probability $1/2$ to each of v_y and v_n, 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 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 choosing an unexpected action. More precisely, suppose that $A_1 = \{a_1, a_1'\}$ and we wish to define u_1 in such a way that player 1 prefers to choose a_1 if and only if player 2 believes he will choose a_1'. In contrast to Example 2, this scenario cannot be represented with a Bayesian game for the following simple reason: *states do not encode actions.* In other words, condition C1 is not satisfied if we take I to be player 1's action. Therefore, types cannot encode such beliefs about actions, so utility cannot be defined in a way that depends on such beliefs.

This suggests an obvious generalization of the Bayesian setting, namely, encoding actions 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 $\beta_i : T_i \to A_i$. In Bayesian games we talk about behaviour rule profiles being in equilibrium, just as in normal-form games, we talk about action profiles being in equilibrium. Intuitively, $\beta_i(t_i)$ represents the action that type t_i of player i is playing, so a player's action 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 action-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 type-signal functions τ_i, a behaviour rule β_i associates with each state ω the action $\beta_i(\tau_i(\omega))$. 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 $p_i(t_i)$. 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}:

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

Finally, a **Bayesian Nash equilibrium** of the Bayesian game $\mathcal{B}$ 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 action and type spaces are finite (see [11] for a more general characterization of when an equilibrium exists).

3. Intention

3.1. Definition

Behaviour rules map types to actions, but it is important to note that the underlying model of a Bayesian game does not enforce any relationship between types and actions (or between states and actions). Thus, behaviour rules cannot do the work required in condition C1, where I consists of the players' actions; thus, they do not allow us to express preferences that depend on beliefs about actions. In order to express such preferences, we must incorporate a connection between actions and types into the game model itself.

Roughly speaking, we accomplish this by expanding the domain of the utility functions to include what we call "intended actions". Formally, an **intention function for player** i is a map $\alpha_i : T_i \to A_i$ associating with each type of player i an action available to player i. Mathematically, of course, intention functions are exactly the same objects as behaviour rules; however, they play conceptually distinct roles in the theory. We have chosen our terminology to reflect that difference. Intuitively, one might think of $\alpha_i(t_i)$ as the action 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" action for that type; it might even be viewed as the "stereotypical" action employed by players of type t_i. The first 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 two interpretations may be appropriate if we think about voting; for example, wealthy people in Connecticut may typically vote Republican, but a particular player i who is wealthy and lives in Connecticut (this information being encoded in her type) might instead vote Democrat.

By a slight abuse of notation, let A^T denote the product $A_1^{T_1} \times \cdots \times A_n^{T_n}$, that is, the set of intention function profiles. The definition of a **Bayesian game with intentions** (BGI) is then exactly the same as the definition of a Bayesian game, with one key modification: the utility functions are taken to be maps $u_i : A \times \Omega \times A^T \to \mathbb{R}$. In other words, in a BGI, players' preferences depend not only on the *actual* actions taken by each player (encoded in the factor A) together with the state of the world (encoded in Ω), but additionally on the *intended* actions of each type of each player (encoded in A^T).

We associate intended actions 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 $\alpha_i \circ \tau_i : \Omega \to A_i$ does associate actions with states and so satisfies condition C1 (again, with I being players' actions). Thus, in this framework, players can have beliefs about actions, and we can define utility so as to capture preferences that depend on such beliefs, as we show by example.

3.2. Examples

The presentation of a BGI is made clearer by introducing the following notation for the set of states where player i intends to play a_i, according to the intention function profile α:

$$[\![a_i]\!]^\alpha = (\alpha_i \circ \tau_i)^{-1}(a_i) = \{\omega \in \Omega \,:\, \alpha_i(\tau_i(\omega)) = a_i\}.$$

Example 3. 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 action ascribe probability 0 to player 1's actual action. Thus, we might define player 1's utility function as follows:

$$u_1(a, \omega, \alpha) = \begin{cases} 1 & \text{if } p_2(\tau_2(\omega))([\![a_1]\!]^\alpha) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

(Recall that $p_2(\tau_2(\omega))$ is a measure on states, which is why we apply it to $\tau_1^{-1}(\alpha_1^{-1}(a_1))$, that is, the set of states ω where player 1's intended action, $\alpha_1(\tau_1(\omega))$, is a_1.) □

Example 4. Next we consider an example introduced by GPS [1] 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 $A_1 = \{\mathsf{bold}, \mathsf{timid}\}$ (and $A_2 = \{*\}$). 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" [1]. It is also stipulated that player 2 prefers player 1 to be bold, and also prefers to think of him as bold. Define $q_\alpha : T \to [0, 1]$ by

$$q_\alpha(t) = p_2(t_2)([\![\mathsf{bold}]\!]^\alpha),$$

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

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

where $E_{t_i}(f)$ denotes the expected value of f with respect to the measure $p_i(t_i)$. Thus $q_\alpha(t)$ captures player 2's assessment of the likelihood that player 1 will be bold, while $\tilde{q}_\alpha(t)$ captures player 1's expectation about player 2's assessment of

that likelihood. We can then represent the players' preferences in a reduced-form BGI as follows:

$$u_1(a,t,\alpha) = \begin{cases} 2 - \tilde{q}_\alpha(t) & \text{if } a_1 = \mathsf{bold} \\ 3(1 - \tilde{q}_\alpha(t)) & \text{if } a_1 = \mathsf{timid}, \end{cases}$$

$$u_2(a,t,\alpha) = \begin{cases} 2(1 + q_\alpha(t)) & \text{if } a_1 = \mathsf{bold} \\ 1 - q_\alpha(t) & \text{if } a_1 = \mathsf{timid}. \end{cases}$$

This representation closely parallels that given in [1], in which q_α and $\tilde{q}_\alpha$ are understood not as functions of types (and intention function profiles), but (implicitly) as functions of belief hierarchies.[5] 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_α is high—that is, all the more so when she believes with high probability that he will be bold.[6] 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 their original paper [1]; specifically, as they define it, a player's utility can depend only on *their own* beliefs. Twenty years later, Battigalli and Dufwenberg [3] corrected this deficiency. BGIs do not encounter such limitations in the first place. In particular, it is easy enough to redefine player 1's utility as follows:

$$u_1'(a,t,\alpha) = \begin{cases} 2 - q_\alpha(t) & \text{if } a_1 = \mathsf{bold} \\ 3(1 - q_\alpha(t)) & \text{if } a_1 = \mathsf{timid}. \end{cases}$$

In this case, we find that player 1 prefers to be timid provided $q_\alpha(t) < \frac{1}{2}$, 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 BGI, in that we did not explicitly define the states, types, 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 BGIs. Particular

[5] 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.

[6] It is not quite clear why GPS define player 2's payoff in the event that player 1 is timid to be $1-q$ rather than $1+q$. 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.

choices of type spaces 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

It is easy to see that given a BGI $\mathcal{I}$ and a profile α of intention functions, we can construct a standard Bayesian game $\mathcal{I}(\alpha)$ by defining new utility functions $\tilde{u}_i : A \times \Omega \to \mathbb{R}$ as follows:

$$\tilde{u}_i(a, \omega) = u_i(a, \omega, \alpha).$$

Call $\mathcal{I}(\alpha)$ an **instantiation of** $\mathcal{I}$. Intuitively, in the Bayesian game $\mathcal{I}(\alpha)$, the profile α gives the "true" intention of each type of each player, and then the preferences of all players are determined with respect to these fixed intentions. This way of producing a standard Bayesian game from a BGI allows us to define a natural notion of equilibrium in our setting.

Say that a profile of behaviour rules β is an **equilibrium of** $\mathcal{I}$ provided β is a Bayesian Nash equilibrium of the instantiated Bayesian game $\mathcal{I}(\beta)$. Here we make implicit use of the fact that both behaviour rules and intention functions are maps from types to actions. Indeed, the profile β plays two roles in this definition. First, it is used to determine the *intended actions* of the players. Then, in the context of the instantiated Bayesian game (with these fixed intentions), we evaluate whether each β_i is a best response, exactly as in the definition of equilibrium for a standard Bayesian game; in this latter role, it is natural to think of β_i as outputting the *actual actions* of the players.

This definition 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 BGI, beliefs about the actions of one's opponents are beliefs about *intended* actions. On the other hand, since behaviour rules associate actions with types and players have beliefs about types, behaviour rules also induce beliefs about actions; in our terminology, these are beliefs about *actual* actions. Our definition of equilibrium requires that these two beliefs coincide: that is, in equilibrium, each type of each player *actually* plays the action she *intends* to play (which is exactly the action her opponents expected (that type of) her to play).

Does this collapse the distinction between intended and actual actions, 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 actions. In addition, although we demand that intended and actual actions 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

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

Crucially, β_i' is permitted to range over *all* behaviour rules. 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 the fixed

profile of intended actions given by β in $\mathcal{I}(\beta)$. Example 5 demonstrates that the notion of best response in an instantiation of a BGI—and therefore the notion of equilibrium—can be sensitive to states of play where players are *not* playing their intended strategies.

Example 5. Consider a 2-player reduced-form BGI $\mathcal{I}$ with $A_1 = \{\mathsf{left}, \mathsf{right}\}$, $A_2 = \{*\}$, $T_1 = \{x, x'\}$, 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, encoding player 1's desire to surprise her opponent:

$$u_1(a, t, \alpha) = \begin{cases} 1 & \text{if } p_2(t_2)([\![a_1]\!]^\alpha) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

First consider an intention function α_1 with $\alpha_1(x) = \alpha_1(x') = \mathsf{left}$ (player 2's intentions are always trivial, so we omit them). Then, of course, $p_2(y)([\![\mathsf{left}]\!]^\alpha) = p_2(y')([\![\mathsf{left}]\!]^\alpha) = 1$, and likewise $p_2(y)([\![\mathsf{right}]\!]^\alpha) = p_2(y')([\![\mathsf{right}]\!]^\alpha) = 0$. It follows immediately that the expected utility of playing left for either type of player 1 is 0 (since player 1 is sure that this will not surprise her opponent), whereas the expected utility of playing right for either type of player 1 is 1 (since, in this case, player 1 is sure that this will surprise her opponent). In particular, if $\beta_1 = \alpha_1$, then β_1 is not a best response in the Bayesian game $\mathcal{I}(\beta)$, so β_1 cannot be part of an equilibrium for $\mathcal{I}$.

Next consider $\alpha_1'(x) = \mathsf{left}$ and $\alpha_1'(x') = \mathsf{right}$. It is not hard to check that if $\beta_1 = \alpha_1'$, then β_1 is a best response in the game $\mathcal{I}(\beta)$, and therefore β is an equilibrium (since, again, player 2's action space is trivial). Indeed, 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 right; thus, left is a best response for x, since x is sure that it will surprise her opponent. A similar argument shows that right is a best response for x'. □

Is this a reasonable notion of equilibrium? Essentially, it encodes the following question: "Is there a profile of intentions such that, assuming those intentions are common knowledge, no player prefers to deviate from their intention?" When the answer is yes, that profile constitutes an equilibrium. At least at a high level, this does seem to accord with a standard notion of what counts as a "stable" state of play. Moreover, as we show in Sections 4 and 5, the notions of equilibrium proposed by GPS for psychological games and by KR [7] for reference-dependent preferences are closely linked with our definition.

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 6. Consider a 2-player reduced-form BGI $\mathcal{I}$ where $A_1 = \{\mathsf{left}, \mathsf{right}\}$, $A_2 = \{*\}$, $T_1 = \{x, x'\}$, 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 5:

$$u_1(a, t, \alpha) = \begin{cases} 1 & \text{if } p_2(t_2)([\![a_1]\!]^\alpha) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Note that the probability map p_2 is not the same as it was in Example 5; in particular, in this game player 1 is certain that player 2 knows her type. It follows that no matter what player 1's 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* β_1 can be a best response in the instantiated Bayesian game $\mathcal{I}(\beta)$; it follows immediately that $\mathcal{I}$ admits no equilibria. □

This obstacle persists even if we extend our attention to mixed strategies. More precisely, consider the class of BGIs where, for each player i, their action set is of the form $\Sigma_i = \Delta(A_i)$ for some finite set A_i (the set of player i's *pure strategies*), and $u_i : \Sigma \times \Omega \times \Sigma^T \to \mathbb{R}$ satisfies

$$u_i(\sigma, \omega, \alpha) = \sum_{a \in A} \Big(\prod_{j \in N} \sigma_j(a_j) \Big) u_i(a, \omega, \alpha).$$

In other words, as usual, the utility of σ is just the expected value of the utility of the various pure strategies with the probabilities induced by σ. As is standard, we call elements of Σ_i *mixed strategies*, and the corresponding BGIs *mixed-strategy BGIs*. We can similarly define *mixed-strategy BGIs*. Note that in this context, since the intention functions α_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 7. Consider a 2-player reduced-form mixed-strategy BGI where $\Sigma_1 = \Delta(\{\mathsf{left}, \mathsf{right}\})$, $\Sigma_2 = \{*\}$, $T_1 = \{x, x'\}$, 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

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

and

$$u_1(\mathsf{right}, *, t, \alpha) = \begin{cases} 1 & \text{if } p_2(t_2)([\![\mathsf{left}]\!]^\alpha) = 1 \\ 0 & \text{otherwise,} \end{cases}$$

and extend to all $\sigma_1 \in \Delta(\{\mathsf{left}, \mathsf{right}\})$ by taking expectation:

$$u_1(\sigma_1, *, t, \alpha) = \sigma_1(\mathsf{left})u_1(\mathsf{left}, *, t, \alpha) + \sigma_1(\mathsf{right})u_1(\mathsf{right}, *, t, \alpha).$$

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

This game admits no equilibria. To see this, suppose that β is an equilibrium. It follows that β_1 is part of a Bayesian Nash equilibrium in the instantiated Bayesian game $\mathcal{I}(\beta)$. First consider the case where $\beta_1(x) \in \Sigma_1$ satisfies $\beta_1(x)(\mathsf{right}) > 0$. Then it follows that $p_2(y)([\![\mathsf{left}]\!]^\beta) = 0$ (i.e., type y is certain that player 1 is not playing the pure strategy left), 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 left. In particular, $\beta_1(x)$ is not a best response, so β_1 cannot be part of a Bayesian Nash equilibrium in $\mathcal{I}(\beta)$. Now consider the case where $\beta_1(x)(\mathsf{right}) = 0$; in other words, $\beta_1(x)$ is the pure strategy left. Then we have $p_2(y)([\![\mathsf{left}]\!]^\beta) = 1$, from which it follows that type x's best response is to play the pure strategy right. Thus, once again, β_1 cannot be part of a Bayesian Nash equilibrium. □

4. Psychological games

Static psychological games can be captured in our framework. A static psychological game $\mathcal{P}$, as defined by GPS, 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 $(b_i^1, b_i^2, \ldots)$, 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 are *collectively coherent* (roughly speaking: lower-order beliefs must agree with the appropriate marginals of higher-order beliefs, and this agreement must be common knowledge). See [1] for the complete definition.

It is well known that each type in a type space encodes a hierarchy of beliefs, so it is perhaps unsurprising that given a mixed-strategy BGI $\mathcal{I}$, a type $t_i \in T_i$, and a profile α of intention functions, we can generate a collectively coherent sequence of beliefs in a natural way. More precisely, define the **first-order beliefs associated with t_i given α (in $\mathcal{I}$)** by

$$\phi_i^1(t_i, \alpha) = (\alpha_{-i} \circ \tau_{-i})_*(p_i(t_i));$$

that is, the pushforward of $p_i(t_i)$ from Ω to Σ_{-i} by $\alpha_{-i} \circ \tau_{-i}$.[7] In other words, for each event $E \subseteq \Sigma_{-i}$, we have

$$\phi_i^1(t_i, \alpha)(E) = p_i(t_i)((\alpha_{-i} \circ \tau_{-i})^{-1}(E)).$$

Note that, in our terminology, $\phi_i^1(t_i, \alpha)$ is a belief about *intended* strategies. The *kth-order beliefs associated with* t_i *given* α, denoted $\phi_i^k(t_i, \alpha)$, can be defined inductively in a similar fashion; it is then straightforward to show that any sequence of the form

$$\phi_i^\alpha(t_i) = (\phi_i^1(t_i, \alpha), \phi_i^2(t_i, \alpha), \ldots)$$

is collectively coherent, and thus $\phi_i^\alpha : T_i \to \bar{B}_i$.

This correspondence provides a natural notion of agreement between a BGI and a static psychological game with respect to the psychological preferences expressed in the latter, namely:

$$\forall i \in N \, \forall \sigma \in \Sigma \, \forall \omega \in \Omega \, \forall \alpha \in \Sigma^T (u_i(\sigma, \omega, \alpha) = v_i(\phi_i^\alpha(\tau_i(\omega)), \sigma)).$$

When a BGI $\mathcal{I}$ satisfies this condition with respect to a static psychological game $\mathcal{P}$, we say that $\mathcal{I}$ **defers to** $\mathcal{P}$. Even very simple BGIs (i.e., those with very small type/state spaces) can defer to static psychological games; it is sufficient that the utility functions u_i be of the form

$$u_i(\sigma, \omega, \alpha) = v_i(\phi_i^\alpha(\tau_i(\omega)), \sigma),$$

so that utility in the BGI depends on states *only to the extent that states encode belief hierarchies.* In particular, although the utility functions in a static psychological game have uncountable domains (since the domain includes all possible belief hierarchies), a BGI $\mathcal{I}$ might defer to $\mathcal{P}$ even if $\mathcal{I}$ has only finitely many states, since all that matters is that the utility functions of $\mathcal{I}$ agree with the utility functions of $\mathcal{P}$ on the belief hierarchies encoded by the states in $\mathcal{I}$.

Given a static psychological game $\mathcal{P}$, it is possible to construct a BGI that defers to $\mathcal{P}$ that also has type spaces rich enough such that each ϕ_i is surjective: in other words, every belief hierarchy is realized by some type. This might seem a natural construction, particularly if the goal is to import static psychological games into the present framework. However, in order to capture the *equilibrium* behaviour of static psychological games, such richness turns out to be superfluous. The notion of equilibrium defined by GPS can in fact be recovered as equilibria in our setting in a much simpler manner.

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 $\mathcal{P}$ is a strategy profile σ such that, for each player i, σ_i maximizes the function

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

[7] Here $\alpha_{-i} \circ \tau_{-i}$ maps each state ω to the corresponding profile of intended strategies for player i's opponents.

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 1. *If $\mathcal{I}$ defers to $\mathcal{P}$, then σ is a psychological Nash equilibrium of $\mathcal{P}$ if and only if the profile of constant behaviour rules β satisfying $\forall t_i(\beta_i(t_i) = \sigma_i)$ is an equilibrium of $\mathcal{I}$.*[8]

Proof. Let β be the profile of behaviour rules described in the theorem. Now β is an equilibrium of $\mathcal{I}$ iff, for each i, β_i is a best response in the instantiated Bayesian game $\mathcal{I}(\beta)$. That is, for each type t_i of player i, playing the strategy recommended by β_i, namely σ_i, must maximize expected utility in $\mathcal{I}(\beta)$. Since $\mathcal{I}$ defers to $\mathcal{P}$, we know that the utility functions $\tilde{u}_i$ in $\mathcal{I}(\beta)$ depend on states only to the extent that states encode belief hierarchies:

$$\tilde{u}_i(\sigma', \omega) = v_i(\phi_i^{\beta}(\tau_i(\omega)), \sigma').$$

Moreover, by definition of β, for each player i we know that every type of player i intends to play σ_j; it follows that this is common knowledge, so for all $i \in N$ and all $\omega \in \Omega$, we have $\phi_i^{\beta}(\tau_i(\omega)) = \chi_i(\sigma)$. This yields

$$\tilde{u}_i(\sigma', \omega) = v_i(\chi_i(\sigma), \sigma').$$

Notice that the term on the right does not mention ω! Thus, $\tilde{u}_i$ is constant in its second component, so to maximize its expected value (where the expectation is taken over Ω) is just to maximize its value, which is precisely the value of $v_i(\chi_i(\sigma), \sigma')$. Thus, β_i is a best response to β_{-i} just in case σ_i maximizes the value of the function

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

It follows that β is a Bayesian Nash equilibrium of $\mathcal{I}(\beta)$ iff each β_i is a best response to β_{-i} iff σ is a psychological Nash equilibrium, which establishes the desired equivalence. □

Theorem 1 shows that equilibrium analysis in static psychological games does not depend on the full space of belief hierarchies; in fact, 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 static 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.

[8] A similar connection between between Bayesian equilibrium and psychological Nash equilibrium was also independently observed by Attanasi et al. [12].

We can view actions as functions of types in dynamic games, and in fact understand Bayesian games as special cases of dynamic games: roughly speaking, nature moves first, choosing the type profile, and the rest of the players act simultaneously afterwards, completing the game. Adapting this translation, we can also view BGIs as special cases of dynamic psychological games in the sense of BD. So, in this sense, BGIs can be viewed as an intermediate point between static psychological games and dynamic psychological games; that is,

$$\{\text{static psych. games}\} \subset \{\text{BGIs}\} \subset \{\text{dynamic psych. games}\}.$$

This is perhaps a natural state of affairs given the high level of generality provided by the dynamic psychological games framework. The value of the BGI framework consists in the range of psychological phenomena it is capable of capturing given its relative simplicity as a modest generalization of the well-known Bayesian games paradigm.

5. Reference-Dependent Preferences

One of the core themes of prospect theory is the idea that preferences depend not only on final outcomes, but also on how those outcomes compare to some "reference point": typically, whether they are evaluated as gains or losses in the context of this comparison. There are many mathematical formalisms one might employ to capture this type of reference dependence, but aside from the question of the mathematical implementation of the comparison, "predictions of reference-dependent theories also depend crucially on the under-studied issue of what the reference point is." [7] Kőszegi and Rabin not only highlight this issue, but also propose a model of reference-dependent preferences that directly addresses it: roughly speaking, they propose that a player's reference point be identified with their beliefs about outcomes.

Bayesian games with intentions seem well-suited to capturing the kinds of comparisons so central to prospect theory. Indeed, it is quite natural to interpret the intended actions in a BGI as establishing a reference point, and define utility in such a way that the subjective value of an actual outcome depends on how it compares to the intended (i.e., the reference) outcome. To make these intuitions precise, we present in this section a reinterpretation of some of KR's formalism and results in the BGI framework; this provides a concrete demonstration of how BGIs can be leveraged to capture key notions in prospect theory. The fact that intended actions in the BGI setting are precisely the actions that players have beliefs about aligns nicely with KR's proposal that reference points are determined by beliefs. Moreover, the fundamental assumption underlying the notion of "personal equilibrium" as defined by KR—namely, that the predictions that constitute a player's reference point are *correct* predictions—is recapitulated in the requirement that equilibria in BGIs demand agreement between intended and actual strategies.

Given "consumption" c and "reference level" r, KR assume that a player's overall utility factors into two parts, "consumption utility" $m(c)$ and "gain-loss

utility" $n(c|r)$:

$$u(c|r) = m(c) + n(c \,|\, r).$$

Moreover, they assume that gain-loss utility depends functionally only on the difference between the real consumption utility and that of the reference level:

$$n(c \,|\, r) = \mu(m(c) - m(r)),$$

where μ is a function satisfying certain conditions that implement features common in prospect theory, such as loss aversion and diminishing sensitivity. For the purposes of this overview, we suppress the details of these conditions.[9]

BGIs are flexible enough to implement an analogous utility structure, identifying consumption with the actual outcome and the reference level with the intended outcome as suggested above. More precisely, say that a BGI $\mathcal{I}$ **implements reference-dependence** if, for each player $i \in N$, there are functions $m_i : A \to \mathbb{R}$ and $n_i : A \times A \to \mathbb{R}$ where, for all $c, r \in A$,

$$n_i(c \,|\, r) = \mu(m_i(c) - m_i(r)),$$

and, for all i,

$$u_i(a, \omega, \alpha) = m_i(a) + n_i(a \,|\, \alpha(\tau(\omega))). \tag{1}$$

Since $\alpha(\tau(\omega))$ is precisely the profile of intended actions played at state ω, we see that this setup treats intended actions exactly as the KR model treats reference points (i.e., as the second input to the gain-loss function n_i). Player i maximizes expected utility by choosing an action that maximizes the expected value of u_i with respect to the beliefs $p_i(t_i)$ determined by i's type t_i. These are beliefs about the state space Ω, and therefore about intended actions; indeed, the form of player i's utility function guarantees that i cares only about the state ω to the extent that it determines the profile of intended actions $\alpha(\tau(\omega))$ to be used as a reference point in the gain-loss function n_i.

Recall that in a BGI, an equilibrium consists of a profile α of intention functions such that each type of each player i maximizes expected utility by actually playing the intended strategy given by α. In the context of a BGI that implements reference-dependence, an equilibrium thus corresponds to a situation where, for each player i, the reference outcomes that i expects (as determined by i's intended actions and the intended actions of i's opponents) induces i to play in such a way that actually establishes those reference outcomes. We believe that this exactly captures KR's gloss of their equilibrium notion: "our notion of *personal equilibrium* assumes that a person correctly predicts both the environment she faces ... and her own reaction to that environment ... and taking the reference point generated by these expectation as given, in each contingency maximizes expected utility." [7] We now turn to an extended example that illustrates this alignment.

[9] We also make the simplifying assumption that consumption is "one-dimensional", rather than a "bundle" of additively separable components $c = (c_1, \ldots, c_K)$, as KR assume.

5.1. Shopping for Shoes

We explore an example 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 follow a parallel path within the BGI framework and compare this approach to that of KR.

We model the scenario as a reduced-form BGI $\mathcal{I}$ with two players $N = \{C, R\}$: a consumer C and a retailer R. As we are interested only in the consumer's decisions and motivations, we capture the retailer's preferences with a constant utility function; in essence, R plays the role of "the environment".

Let A_R be a set of non-negative real numbers that represent possible prices of a pair of shoes. The retailer must choose a price $p \in A_R$. 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 options available to C should reflect this. Let A_C be a set of real numbers, the *thresholds*; $\theta \in A_C$ represents the threshold cost at which C is no longer willing to buy the shoes. An outcome is therefore a threshold-price pair $(\theta, p) \in A$; intuitively, the shoes are purchased for price p if and only if $p < \theta$. For convenience we assume that both A_R and A_C are finite.

The consumer's utility depends on the outcome of the game together with the reference level; more precisely, the two dimensions of consumption utility are given by functions $m_i : A \to \mathbb{R}$ defined by

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

and

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

As KR do, we assume *additive separability* of consumption utility, so the function $m_C = m_1 + m_2$ 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. We can thus think of a price $p > 1$ as being "overpriced" in the sense that purchasing the shoes for this price results in a consumption utility lower than not purchasing the shoes at all.

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

$$n_i(\theta, p \,|\, \theta', p') = \mu(m_i(\theta, p) - m_i(\theta', p')),$$

where, following KR, we set

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

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 $n_C = n_1 + n_2$ gives the total gain-loss utility. Finally, C's total utility is given by the sum of her consumption utility m_C and her gain-loss utility n_C, as in Equation (1):

$$u_C(\theta, p, t, \alpha) = m_C(\theta, p) + n_C(\theta, p \,|\, \alpha_C(t_C), \alpha_R(t_R)).$$

We begin, as KR do, by considering the consumer's behaviour in relatively simple scenarios of price certainty (i.e., when the consumer is certain of the price the shoes will be offered at). KR show that, in this case, both buying for sure and not buying for sure can be personal equilibria for the consumer, provided the price is not *too* high or low. This result 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 might pass up a good deal if they had their mind set in advance on saving their money.

KR work in a dynamic setting where these intuitions can be cashed out temporally: first, the consumer forms an expectation that she will buy the shoes before she even gets to the store; then, upon arrival, she realizes (say) that they are more expensive than she had thought, and updates her beliefs accordingly. Crucially, however, she does not update her reference level—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.)

The BGI framework we have developed does not include an explicit temporal component. Nonetheless, as we have emphasized, the distinction between intended and actual actions provides a mechanism that can be leveraged to capture the very same sensations of gain and loss described above. Moreover, since intended actions are what players have direct beliefs about, this accords with the intuition that the consumer's reference level is determined by her expectations—namely, what she believes the price might be, and whether she intends to buy at that price. We show that the equilibria in our framework correspond exactly to the "personal equilibria" in KR's setting.

Recall that a behaviour rule β constitutes an equilibrium provided β is a Bayesian Nash equilibrium of the corresponding Bayesian game $\mathcal{I}(\beta)$. Since R has a flat utility function, this amounts to requiring that each type t_C of C is "rational", that is, that $\beta_C(t_C)$ maximizes expected utility according to $p_C(t_C)$.

To capture the case of price certainty, we set $A_R = \{p\}$. Suppose first that $\beta_C(t_C) = \theta'$ and $p < \theta'$; in other words, type t_C of C intends to buy the shoes.

Since t_C is certain about both the price of the shoes (p) and her intention to buy them at that price (θ'), the expected utility computation is particularly simple: choosing $\theta \in A_C$ yields (in all cases) utility

$$m_C(\theta, p) + n_C(\theta, p \,|\, \theta', p).$$

Consider $\theta_L, \theta_H \in A_C$ with $\theta_L < p < \theta_H$. It is easy to calculate that

$$\begin{aligned} m_C(\theta_L, p) + n_C(\theta_L, p \,|\, \theta', p) &= 0 + \mu(0 - (-p)) + \mu(0 - 1) \\ &= \eta p - \lambda\eta \end{aligned}$$

and

$$\begin{aligned} m_C(\theta_H, p) + n_C(\theta_H, p \,|\, \theta', p) &= 1 - p + \mu(-p - (-p)) + \mu(1 - 1) \\ &= 1 - p; \end{aligned}$$

therefore, C can rationally choose θ_H rather than θ_L whenever

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

Note that this does not depend on the actual values of θ_L and θ_H, but only their relation to p. In particular, since by supposition $\beta_C(t_C) = \theta' > p$, we see that C maximizes expected utility by following through on her intentions and playing θ' just in case the above inequality holds. Since the righthand side of this inequality is greater than 1, this shows that intending to buy can make it rational to actually buy, even for some prices $p > 1$, i.e., when the shoes are "overpriced".

Similarly, if $\beta_C(t_C) = \theta'' < p$, analogous calculations yield

$$m_C(\theta_L, p) + n_C(\theta_L, p \,|\, \theta'', p) = 0$$

and

$$m_C(\theta_H, p) + n_C(\theta_H, p \,|\, \theta'', p) = 1 - p - \lambda\eta p + \eta,$$

which shows that C is maximizing expected utility whenever

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

so intending not to buy makes it rational not to buy even for some prices $p < 1$, i.e., when the shoes are "underpriced". These findings duplicate those of KR.

Next we consider the case of price *uncertainty* presented by KR. Let $A_R = \{p_L, p_M, p_H\}$, where $p_L < p_M < p_H$, and suppose that t_C is a type for which

$$p_C(t_C)([\![p_L]\!]^\alpha) = q_L$$

and

$$p_C(t_C)([\![p_H]\!]^\alpha) = q_H = 1 - q_L,$$

for some profile of intention functions α. Thus, the consumer ascribes probability 0 to the price being p_M in the context of α. Suppose also, for simplicity, that $A_C = \{\theta_L, \theta_H\}$, where $p_L < \theta_L < p_M < \theta_H < p_H$. Thus, the two actions available to C constitute a choice between buying at price p_M or not, while buying at price p_L is a foregone conclusion and buying at price p_H is off the table entirely.

Following KR, we will "examine the consumer's willingness to pay without worrying about how her behaviour feeds back into her expectations ... and consider the 'out-of-equilibrium' question of whether she buys at the intermediate price p_M." [7] Of course, the meaning of "out-of-equilibrium" differs in our two settings. At a high level, the issue is that we are considering a consumer who does not expect the price p_M (her beliefs are split between p_L and p_H), yet nonetheless finds herself faced with that price. In KR's setting, this is "out-of-equilibrium" because the reference expectations are not correct. In our framework, this corresponds to a case where intended actions, and therefore also beliefs about intended actions, are inconsistent with actual actions.

In the Bayesian game $\mathcal{I}(\alpha)$ instantiated by α, instead of asking whether α itself is a Bayesian Nash equilibrium (as required by our definition of equilibria in BGIs), we instead focus on a behaviour rule β for which β_R always outputs p_M (i.e., the retailer actually offers the price p_M), and ask what constitutes a best response for type t_C of player C in this case. So, in fact, we are still "in-equilibrium" in a certain sense, namely, that we are searching for a Bayesian Nash equilibrium of $\mathcal{I}(\alpha)$; yet we are also "out-of-equilibrium" in the sense that we have dropped the requirement that intended and actual actions coincide. In short, the beliefs of t_C about intended actions, which determine her reference point, differ from her beliefs about actual actions, with respect to which she must maximize expected utility.

The expected utility for t_C is given by:

$$q_L \cdot u_C(\beta_c(t_C), p_M \,|\, \alpha_C(t_C), p_L) + q_H \cdot u_C(\beta_C(t_C), p_M \,|\, \alpha_C(t_C), p_H).$$

Notice that $p_L < \alpha_C(t_C) < p_H$, so in fact the precise value of $\alpha_C(t_C)$ is irrelevant (the consumer always buys at price p_L and abstains at price p_H, which fix her reference points accordingly). Thus, when $\beta_C(t_C) = \theta_H$, the expected utility is

$$\begin{aligned} & q_L(1 - p_M + \mu(-p_M - (-p_L))) + \mu(0)) + q_H(1 - p_M + \mu(-p_M) + \mu(1)) \\ = \;& 1 - p_M + q_L\lambda\eta(p_L - p_M) + q_H(\lambda\eta(-p_M) + \eta), \end{aligned}$$

while the expected utility when $\beta_C(t_C) = \theta_L$ is

$$\begin{aligned} & q_L(\mu(0 - (-p_L)) + \mu(-1)) + q_H(\mu(0) + \mu(0)) \\ = \;& q_L(\eta p_L - \lambda\eta). \end{aligned}$$

These are precisely the values that KR calculate, and so the lessons they draw in their setting are equally available in ours. First, the "attachment effect":

when $p_L = 0$ (so the consumer believes the shoes may be available for free), an increase in q_L (which increases the expectation to get them for free) also increases the sense of loss felt if the shoes are not, ultimately, purchased (for the true price of p_M), which in turn increases the price the consumer is actually willing to pay). And second, the "comparison effect": when $p_L \geq 0$ and $q_L = 1$ (so the consumer thinks for sure the shoes will be available at the low price p_L), a decrease in p_L increases the sense of loss felt by buying the shoes at the higher price of p_M, and therefore decreases her willingness to pay that higher price.

6. Conclusion

We have introduced Bayesian games with intentions, generalizing Bayesian games and static psychological games in a natural way. Whereas psychological games represent players' beliefs using hierarchies of probability measures, BGIs instead deploy the familiar types formalism, which we feel helps to broaden the appeal and accessibility of the psychological game theory paradigm. Moreover, under our translation, psychological Nash equilibria are revealed to be, in essence, special kinds of Bayesian Nash equilibria, further strengthening this connection and its usefulness.

We've also shown that the distinction between real and reference outcomes so important in prospect theory can be represented naturally using actual and intended outcomes, respectively, allowing us to capture many of the key features of prospect theory in the BGI framework. In particular, the models for reference-dependence proposed by Kőszegi and Rabin can be implemented as BGIs in a way that preserves the core insights of their theory, both in and out of equilibrium.

Although BGIs have their roots in Bayesian games and have many features in common with them, we have seen that the addition of "intention", and the corresponding new notion of equilibrium, changes the applicability of this framework substantially. Many important questions remain to be worked out, both theoretical and practical.

On the theoretical side, most notably: when do equilibria exist? While Theorem 1 provides sufficient conditions for the existence of equilibria in BGIs, they are certainly not necessary conditions. It can be shown, 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 an interesting direction for future work.

On the more practical side, we have seen that BGIs can implement a version of reference-dependence, and also that they are expressive enough to capture static psychological games, but what about both at once? There are many potential avenues to explore along these lines; we highlight just one. In our formalization of reference-dependence, we assumed that the gain-loss utility associated with a state ω is a function of the actual and intended outcome *at that state*. However, one might plausibly argue that, insofar as the intended outcome is taken to determine a reference point, and reference points are purely

subjective entities, the reference point should instead be determined by the player's *expectations* at ω.

Concretely, for finite action spaces, we might define

$$u_i(a, \omega, \alpha) = m_i(a) + \sum_{r \in A} \Big(n_i(a|r) \cdot p_i(\tau_i(\omega))([\![r]\!]^\alpha) \Big).^{10}$$

In other words, player i's utility is the sum of two values: (1) i's consumption utility $m_i(a)$, and (2) the *expected* value of i's gain-loss utility, where the expectation is taken with respect to i's beliefs at ω. Since this definition makes i's beliefs directly relevant to her preferences, it is, intuitively, a kind of psychological game. On the other hand, it's also clearly an attempt to implement a certain kind of reference-dependence. We leave the investigation of this class of games, as well as other games definable in the BGI framework, to future work.

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, and Pierpaolo Battigalli for extremely detailed and thoughtful comments on an earlier version of this paper. Halpern was supported in part by NSF grants IIS-0911036, IIS-1718108, IIS-1703846, and CCF-1214844, by ARO grant W911NF-17-1-0952, the Open Philanthropy Foundation, 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, NSF grant IIS-1703846, AFOSR Award FA9550-08-1-0197, and BSF Grant 2006317.

References

[1] J. Geanakoplos, D. Pearce, E. Stacchetti, Psychological games and sequential rationality, Games and Economic Behavior 1 (1) (1989) 60–80. doi:10.1016/0899-8256(89)90005-5.

[2] J. Harsanyi, Games with incomplete information played by 'Bayesian' players, parts I–III, Management Science 14 (1968) 159–182, 320–334, 486–502. doi:10.1287/mnsc.14.5.320.

[3] P. Battigalli, M. Dufwenberg, Dynamic psychological games, Journal of Economic Theory 144 (2009) 1–35. doi:10.1016/j.jet.2008.01.004.

[4] P. Battigalli, R. Carrao, M. Dufwenberg, Incorporating belief-dependent motivation in games, iGIER working paper 642, Bocconi University (2019).

[5] P. Battigalli, M. Dufwenberg, A. Smith, Frustration and anger in games, iGIER working paper 539, Bocconi University (2015).

[10] Naturally, we define $[\![r]\!]^\alpha = \bigcap_{i=1}^n [\![r_i]\!]^\alpha$.

[6] D. Kahneman, A. Tversky, Prospect theory, an analysis of decision under risk, Econometrica 47 (2) (1979) 263–292. doi:10.2307/1914185.

[7] B. Kőszegi, M. Rabin, A model of reference-dependent preferences, The Quarterly Journal of Economics CXXI (2006) 1133–1165. doi:10.1093/qje/121.4.1133.

[8] M. J. Osborne, A. Rubinstein, A Course in Game Theory, MIT Press, Cambridge, MA, 1994.

[9] D. Fudenberg, J. Tirole, Game Theory, MIT Press, Cambridge, MA, 1991.

[10] E. Dekel, M. Siniscalchi, Epistemic game theory, in: H. P. Young, S. Zamir (Eds.), Handbook of Game Theory with Economic Applications, Volume 4, North-Holland, Amsterdam, 2015, pp. 619–702.

[11] G. Tian, The existence of equilibria in games with arbitrary strategy spaces and payoffs: a full characterization (2009).
URL `http://www.dklevine.com/archive/refs4814577000000000160.pdf`

[12] G. Attanasi, P. Battigalli, E. Manzoni, Incomplete information models of guilt aversion in the trust grame, Management Science 62 (2016) 648–667.