

# Dynamic Psychological Games\*

Pierpaolo Battigalli

Martin Dufwenberg

Bocconi University

University of Arizona

September 21, 2005

## Abstract

The motivation of decision makers who care for emotions, reciprocity, or social conformity may depend directly on beliefs (about choices, beliefs, or information). Geanakoplos, Pearce & Stacchetti (*Games and Economic Behavior*, 1989) point out that traditional game theory is ill-equipped to address such matters, and they pioneer a new framework which does. However, their toolbox – psychological game theory – incorporates several restrictions that rule out plausible forms of belief-dependent motivation. Building on recent work on dynamic interactive epistemology, we propose a more general framework. Updated higher-order beliefs, beliefs of others, and plans of action may influence motivation, and we can capture dynamic psychological effects (such as sequential reciprocity, psychological forward induction, regret, and anxiety) that were previously ruled out. We develop solution concepts, provide examples, and explore properties.

KEYWORDS: psychological games, belief-dependent motivation, extensive-form solution concepts, dynamic interactive epistemology.

*J.E.L.* CLASSIFICATION NUMBERS: C72, C73.

---

\*We thank Geir Asheim, Oliver Board, Adam Brandenburger, Amanda Friedenberg, Drew Fudenberg, Georg Kirchsteiger, David Pearce, Klaus Ritzberger, Joel Sobel, and participants in several seminars for helpful discussions. For their kind hospitality, we thank the Economics Departments of the Stern School of Business at NYU, the European University Institute (Battigalli), and Göteborg University (Dufwenberg). Pierpaolo Battigalli gratefully acknowledges financial support from Bocconi University.

# 1 Introduction

We develop a framework for analyzing strategic interaction when players have ‘belief-dependent’ motivations, generalizing the theory of extensive form *psychological games* proposed by Geanakoplos, Pearce & Stacchetti (1989; henceforth GPS). The rest of this introduction motivates in more detail.

Traditional game theory is not a rich enough toolbox to adequately describe many psychological or social aspects of motivation and behavior. The traditional approach assumes utilities depend only on which actions are chosen, but if decision makers are emotional or motivated by reciprocity or social respect utilities may depend also on which beliefs (about choices, beliefs, or information) players harbor. The following examples illustrate:

1. When Ann takes a taxi ride she tips as much as she expects that the driver (Bob) expects to get. She suffers from guilt if she tips less.
2. Cleo suddenly pushes Dan over. Should Dan splash a bucket of water over Cleo in return? Maybe she actually tried to hug him? If so, Dan would rather forgive (maybe even hug) Cleo.
3. Eva is unemployed. Her neighbor, Fred, observes the effort with which she tries to get a job. Fred’s taxes pay for Eva’s unemployment benefits, so Eva’s choice has externalities the size of which depends on her talent translating effort to probability of getting a job (low effort is costlier to Fred if Eva is talented and could have gotten a job had she tried harder). Eva’s talent is known only to her, but Fred makes inferences observing her effort. This determines the social respect he bestows on Eva, and since she cares about respect this influences her effort.

Ann’s tip, Dan’s hug/soak choice, and Eva’s effort each pins down a strategy profile. Yet the preferred choice depends on a belief.<sup>1</sup>

The point that belief-dependent motivation may be important for strategic decision making is made by GPS, who present several intriguing examples involving various emotions. They show the inadequacy of traditional methods to represent the involved preferences, and develop an extension (in the normal as well as in

---

<sup>1</sup>Ann’s preference depends on her belief of Bob’s belief; Dan’s on his assessment of Cleo’s intentions; Eva’s preferences over effort depend on Fred’s inferences on her talent.

the extensive form) of traditional game theory to deal with the matter.<sup>2</sup> Only recently, however, has a larger set of economists come to acknowledge the relevance of belief-dependent motivation, mainly following the work by experimentalists.<sup>3</sup> In the lab, subjects often display ‘non-selfish’ behavior, and this has inspired theoretical models of ‘social preferences’ which can rationalize the data.<sup>4</sup> These models differ in structure, and some do not require a deviation from traditional game theory (*e.g.*, inequity aversion models). However, a few models describe belief-dependent motivation, and some experiments support such models.<sup>5</sup>

While GPS’ paper is highly inspiring for all this work, a careful scrutiny reveals that their approach is too restrictive to handle many plausible forms of belief-dependent motivation (this is acknowledged by GPS themselves; see pp. 70, 78-79). There are several reasons:

**R1 (updated beliefs):** GPS only allow *initial* beliefs to enter the domain of a player’s utility, while many seemingly important forms of belief-dependent motivation require *updated* beliefs to matter.

**R2 (others’ beliefs):** GPS only allow a player’s *own* beliefs to enter the domain of his utility function, while there are conceptual and technical reasons to let *others’* beliefs matter.

**R3 (dependence on strategies):** GPS follow the traditional extensive games approach of letting strategies influence utilities only insofar as they influence terminal histories, but many forms of belief-dependent motivation become compelling in particular in conjunction with preferences that depend on strategies in ways not captured by terminal histories.

**R4 (non-equilibrium analysis):** GPS restrict attention to equilibrium analysis, but in many strategic situations there is little compelling reason to ex-

---

<sup>2</sup>Before GPS, Gilboa & Schmeidler (1988) considered some games with belief-dependent payoffs. Within a decision-theoretic context, Robin Pope has written extensively, over many years, about how conventional theory excludes various forms of belief-dependent motivation. We refer to Pope (2004), which expounds her program and gives further references. Caplin & Leahy (2001) develop a theory of decision making whereby the agent’s utility in each period is a function of his current psychological state, which in turn depends on his beliefs about future outcomes. In later work they relate this theory to psychological games (*e.g.*, Caplin & Leahy, 2004).

<sup>3</sup>See, however, the applied work by Huang & Wu (1994), Dufwenberg (1995, 2002), Geanakoplos (1996), Ruffle (1999), Huck & Kübler (2000), Caplin & Eliaz (2003), Caplin & Leahy (2004), and Li (2005), as well as Bernheim (1994) and Dufwenberg & Lundholm (2000) which can be given psychological-game interpretations (as explained further below).

<sup>4</sup>See Fehr & Gächter (2000) for a discussion.

<sup>5</sup>For models, see Rabin (1993), Dufwenberg & Kirchsteiger (2004), Falk & Fischbacher (2005), Charness & Dufwenberg (2004); for experiments, Dufwenberg & Gneezy (2000), Bacharach, Guerra & Zizzo (2002), Guerra & Zizzo (2004), Charness & Dufwenberg (2004).

pect players to coordinate on an equilibrium and one may wish to explore alternative assumptions.

This list deserves backup by examples, but we postpone this until the next section. Here we just note that items in the list have lead some researchers to deviate from GPS' framework, in developing specific examples or models with belief-dependent motivation. However almost no papers are concerned with developing the overall framework of psychological game theory.<sup>6</sup> We attempt to fill this gap, using **R1-R4** as guiding principles.

Our approach crucially draws on Battigalli & Siniscalchi's (1999) work on how to represent hierarchies of conditional beliefs. This is essential for **R1**, and figures in the background of **R2-R4** which are all related to updated beliefs. We define a large class of psychological games, which contains (in a particular sense) GPS' games and traditional games as special cases. Our main goal is to develop this basic framework, and to illustrate some solution concepts that can be meaningfully developed for it. While one could imagine a variety of interesting solution concepts, we choose to extend two basic concepts of classical game theory to our setting: sequential equilibrium and (extensive form) rationalizability. We prove related theorems, and illustrate how the concepts work in examples.

One final reflection before we proceed: Analyzing games with belief-dependent motivations is intellectually stimulating, because subtle and intriguing conclusions arise. The topic is important, because belief-dependent motivations accords well with (we think) both introspection and experimental evidence. This does not mean that our model can capture all aspects of human motivation. For example, our framework is not (at least primarily) intended to model situations where the preferences of one player depend on the preferences of another. Levine (1998) and, in a more general setup, Gul & Pesendorfer (2005), develop interesting such models of 'interdependent preferences'. Those models do not feature belief-dependent motivation. Theirs and our approaches are best seen as complementary.

Section 2 surveys conceptual issues. Section 3 develops the general framework, up to the definition of a psychological game. Section 4 concerns sequential equilibrium. Section 5 concerns interactive epistemology and rationalizability. Section 6 discusses extensions thus far not covered, including incomplete information. Section 7 concludes. An appendix collects some of the proofs.

---

<sup>6</sup>Kolpin (1992) explores an alternative route to GPS' games, where players 'choose beliefs'. Segal & Sobel (2003) analyze simultaneous move games, and assume preferences over material consequences depend on the equilibrium probability distribution over actions. They show that their approach can be regarded as a reformulation of GPS' normal form games.

## 2 Overview of the conceptual issues

This section surveys the conceptual issues that motivate us. We first describe what GPS' do, and why this is 'non-standard' *vis-a-vis* traditional game theory including models of incomplete information (2.1). We then explain what is our own contribution, going through **R1-R4** from the introduction in more detail (2.2). The style is 'semi-technical', we introduce some notation, but postpone a proper treatment of details for later.

### 2.1 What GPS do

The traditional approach to analyzing extensive games (with complete information) describes a player's preferences using a utility function of the form

$$u_i : Z \rightarrow \mathbb{R}$$

where  $Z$  is the set of terminal histories (endnodes).

Psychological games capture richer motivations than traditional games, and the payoff functions have richer domains. GPS define a set of  $i$ 's initial (pre-play) beliefs about others' strategies and initial beliefs, here referred to as  $\overline{\mathbf{M}}_i$ , which does not rule out any hierarchy of initial beliefs. Each element of  $\overline{\mathbf{M}}_i$  is a sequence  $\overline{\mu}_i = (\overline{\mu}_i^1, \overline{\mu}_i^2, \dots)$  where  $\overline{\mu}_i^1$  represents  $i$ 's beliefs about the opponents' strategies, or first-order beliefs,  $\overline{\mu}_i^2$  represents  $i$ 's joint beliefs about the opponents' strategies and first-order beliefs, and so on.<sup>7</sup>

GPS model preferences using utility functions of the form

$$u_i : Z \times \overline{\mathbf{M}}_i \rightarrow \mathbb{R}$$

This structure bears some superficial similarities to games of incomplete information. It is worth clarifying the differences. In a game of incomplete information some payoff-relevant exogenous parameters (*e.g.* players' abilities or tastes) are not commonly known. Let  $\theta \in \Theta$  denote the vector of such parameters. Note that  $\theta$  does *not* specify strategic choices. A player has beliefs about  $\theta$  (comprising her private information about  $\theta$ ), beliefs about the beliefs of others concerning  $\theta$ , etc. Following Harsanyi (1967-68), such first- and higher-order beliefs can

---

<sup>7</sup>More formally, first-order beliefs are elements of  $\Delta(S_{-i})$ , second-order beliefs elements of  $\Delta(S_{-i} \times \prod_{j \neq i} \Delta(S_{-j}))$ . Upper-bars distinguish initial beliefs from systems of conditional beliefs, the main object of our analysis. We will be more precise in section 3.

be represented in an elegant, albeit implicit, form by assuming that each player  $i$  is characterized by a ‘type’  $t_i \in T_i$  and each  $t_i$  corresponds to a probability measure  $p_{t_i}$  over the set of payoff-relevant parameters and opponents’ types, *i.e.*  $p_{t_i} \in \Delta(\Theta \times T_{-i})$ . It can be shown that  $p_{t_i}$  corresponds to an infinite hierarchy of beliefs  $(p_{t_i}^1, p_{t_i}^2, \dots)$  where  $p_{t_i}^1 \in \Delta(\Theta)$  is the marginal of  $p_{t_i}$  on  $\Theta$ ,  $p_{t_i}^2$  is a joint belief about  $\theta$  and the opponents’ beliefs about  $\theta$ , and so on. The payoff functions of the incomplete information game can be represented as  $V_i : Z \times \prod_j T_j \rightarrow \mathbb{R}$ .

Thus, both psychological games and incomplete information games can be regarded as situations where payoffs depends not only on how the game is played ( $z \in Z$ ) but also on hierarchical beliefs. However, we are talking about different beliefs in the two cases. In psychological games payoffs at endnodes depend on beliefs about strategies, beliefs about such beliefs, and so on. The modeler explains/predicts, such beliefs *via* some solution concept. Hence payoffs at a given endnode are *endogenous*. On the other hand, players’ hierarchical beliefs about the parameter vector  $\theta$  are as *exogenous* as  $\theta$  itself. Hence payoffs at a given terminal history of an incomplete information game are exogenous as well.<sup>8</sup>

## 2.2 Extension of GPS

GPS’ approach can capture interesting forms of belief-dependent motivation. Example 1 of the Introduction, *e.g.*, could be handled by assuming that Ann’s utility equals  $w - m - 2|\bar{\mu} - m|$ , where  $w$  is her pre-tip wealth,  $m \in \{0, 1, \dots, w\}$  her tip, and  $\bar{\mu}$  her expectation of Bob’s expectation of  $m$ , a function of her second-order belief. Ann maximizes her utility by choosing  $m = \bar{\mu}$ .

However, the issues **R1-R4** lead us to enrich the domain of utilities further. We consider payoff functions of the form

$$u_i : Z \times \mathbf{M}_i \times \prod_{j \neq i} (\mathbf{M}_j \times S_j) \rightarrow \mathbb{R}$$

where  $\mathbf{M}_j$  (with  $j = i$  or  $j \neq i$ ) is the set of  $j$ ’s possible *conditional* beliefs about others’ strategies and conditional beliefs,  $S_j$  is the set of (pure) strategies of  $j$ , and  $N$  is the set of players. The conditioning in  $\mathbf{M}_j$  is done for every history, building on Battigalli & Siniscalchi (1999) who show how to represent hierarchies

---

<sup>8</sup>Furthermore, the functional form  $V_i : Z \times \prod_j T_j \rightarrow \mathbb{R}$  is somewhat spurious. The structural payoff functions of an incomplete information game have the form  $v_i : Z \times \Theta \rightarrow \mathbb{R}$ . Function  $V_i$  is obtained as follows:  $V_i(z, t_i, t_{-i}) = \int v_i(z, \theta) p_{t_i}(d\theta | t_{-i})$ . In subsection 6.2, we briefly analyze psychological games with incomplete information. In such games players payoff may directly depend on exogenous as well as endogenous beliefs.

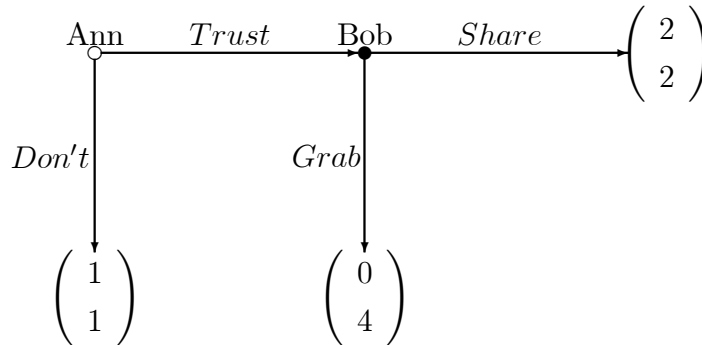
of conditional beliefs without ruling out any hierarchy.  $\overline{\mathbf{M}}_j$  is (isomorphic to) a subspace of  $\mathbf{M}_j$ , so the payoff functions we consider are more general than those assumed by GPS.<sup>9</sup>

Issues **R1-R4** will be related to different arguments of  $u_i$  as we go.

### **R1: updated beliefs**

Rabin’s (1993) theory of reciprocity, in which players reciprocate belief-dependent (un)kindness with (un)kindness, is the most well-known application of GPS’ theory. Rabin works in the normal form. His goal is to highlight key qualitative features of reciprocity, and he does not address issues of dynamic decision making although he points out that this is important for applied work (p. 1296). Dufwenberg & Kirchsteiger (2004) pick up from there, and develop a reciprocity theory for extensive games. In motivating their exercise, they argue that it is necessary to deviate from GPS’ extensive form framework: GPS only allow initial beliefs to enter the domain of a player’s utility, while the modeling of reciprocal response at various ventures of a game tree requires that kindness be re-evaluated using updated belief. The argument is an instance of **R1**.

Reciprocity theory does not provide the easiest route to illustrating the key issues involved though. Instead, we consider the motivation of guilt aversion, applied to the trust game  $\Gamma_1$ .<sup>10</sup> Payoffs are in dollars and do not necessarily represent preferences. Therefore, we call them ‘material payoffs’.



**Figure 1.** The Trust Game  $\Gamma_1$  with material payoffs.

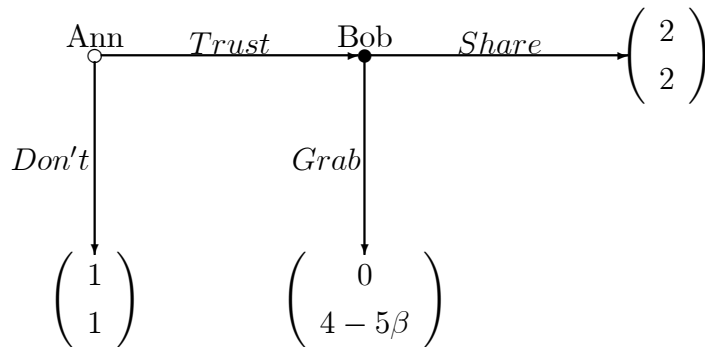
We now modify  $\Gamma_1$  to incorporate a guilt sentiment of Bob’s: To make our point, let us first specify what it means that ‘Bob lets Ann down’. Ann is let down if the material payoff she gets is less than what she expected. Let  $\alpha$  be the probability that Ann (initially) assigns to Bob’s strategy *Share if Trust*. Bob

<sup>9</sup>For a more precise comparison between our framework and GPS see subsection 6.1.

<sup>10</sup>Charness & Dufwenberg’s (2004), coin the term “guilt aversion” and develop a theory within the framework of GPS. Huang & Wu (1994), Dufwenberg (1995, 2002), Dufwenberg & Gneezy (2000), Bacharach, Guerra & Zizzo (2002), and Guerra & Zizzo (2004) consider related sentiments in trust games.

suffers from guilt to the extent that he believes he lets Ann down. He argues that the higher is  $\alpha$  the more let down she will be if he chooses *Grab*. Bob does not know what  $\alpha$  is, as this belief is in the mind of Ann. However, he has a belief about  $\alpha$ . Let  $\beta$  be Bob's expectation of  $\alpha$ , conditional on Ann choosing *Trust*. We can model guilt aversion assuming that Bob's utility at the terminal history (*Trust*, *Grab*) is decreasing in  $\beta$ .

The psychological game  $\Gamma_2$  models this. What appears at the terminal histories should be thought of as utilities, not material payoffs although the notions coincide for all but one terminal histories.<sup>11</sup>



**Figure 2.** The Psychological Trust Game  $\Gamma_2$ .

$\Gamma_2$  is not a psychological game in GPS' class, because  $\beta$  (being an updated belief) is not captured by any element of  $\overline{\mathbf{M}}_i$ . This in itself illustrates **R1**. However, in order to appreciate the significance of this issue, it is useful to note that one can draw compelling (we think) conclusions about behavior that hinge crucially on the fact that  $\beta$  is an updated belief.

Following Dufwenberg (1995, 2002), consider the following (for the time being intuitive) 'psychological forward induction' argument: Suppose Ann chooses *Trust*. If she is rational, she must believe the probability that Bob would choose *Share* (after *Trust*) is at least  $\frac{1}{2}$ , *i.e.*,  $\alpha \geq \frac{1}{2}$ . Since we can figure this out, presumably Bob can too. Even if he is uncertain regarding the value of  $\alpha$ , he infers it is at least  $\frac{1}{2}$ . Hence  $\beta \geq \frac{1}{2}$ . Since  $4 - 5\beta < 2$  if  $\beta \geq \frac{1}{2}$ , he prefers *Share*. Since we can figure this out, presumably Ann can too. Hence she chooses *Trust*, fully expecting Bob to *Share* (so  $\alpha = 1$ ). Bob figures this out (so that  $\beta = 1$ ), which further reinforces his preference to *Share*. The path (*Trust*, *Share*) is predicted!

The logic of the argument depends on belief  $\beta$  being *conditional* on Ann choosing *Trust*. The argument cannot be recast using GPS' theory, since  $\overline{\mathbf{M}}_i$  contains only initial beliefs, but it can be captured in our framework, since  $\mathbf{M}_i$  contains all

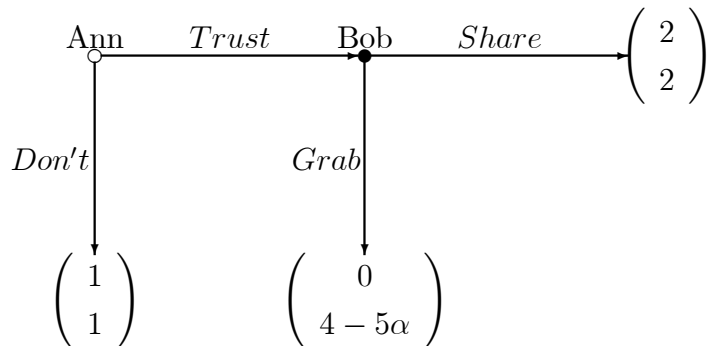
<sup>11</sup>There is no special significance to the "5" in Figure 2; we could have chosen many other numbers to make the upcoming point. Similar remarks apply to all examples below.

hierarchies of conditional beliefs.

## R2: others' beliefs

There are two independent justifications for letting a player's utility depend on others' beliefs. First, this may be an adequate description of how certain social rewards operate. Refer back to ex. 3 from the introduction, where Eva's preferences over effort depends on Fred's inferences. It is taken from Dufwenberg & Lundholm (2001). A related example is Bernheim's (1994) model of social conformity. Another example is Caplin & Leahy's (2004) story of an information providing doctor concerned about the belief-dependent anxiety of a patient.<sup>12</sup> These authors develop models where a player's utility depends on others' beliefs (although only Caplin & Leahy explicitly refer to psychological games).<sup>13</sup>

The second justification concerns convenience in modeling. Refer back to the discussion concerning  $\Gamma_2$ , including the definition of  $\alpha$  and  $\beta$ . In  $\Gamma_2$  we modeled Bob's guilt feelings by letting his psychological payoff depend on  $\beta$ . It turns out that one has an equivalent modeling choice. One can assume that Bob's utility at  $(\ell, L)$  depends directly on  $\alpha$ , rather than on  $\beta$ , although Bob is uncertain about the true value of  $\alpha$  so that he uses probability assessments to weigh the different possibilities. We then get  $\Gamma_3$ :



**Figure 3.** The Psychological Trust Game  $\Gamma_3$ .

After *Trust*, when Bob has to make a choice he compares 2, the payoff of action *Share*, with the conditional expected payoff of action *Grab*, that is  $E_2[4 - 5\alpha | \text{Trust}] = 4 - 5\beta$ ; thus, we obtain the same results as with  $\Gamma_2$ .<sup>14</sup>

This example illustrates an important point: some belief-dependent motivations can be modeled replacing a conditional own belief of a certain 'order' (mean-

<sup>12</sup>For related work emphasizing policy issues see Caplin (2003), Caplin & Eliaz (2003).

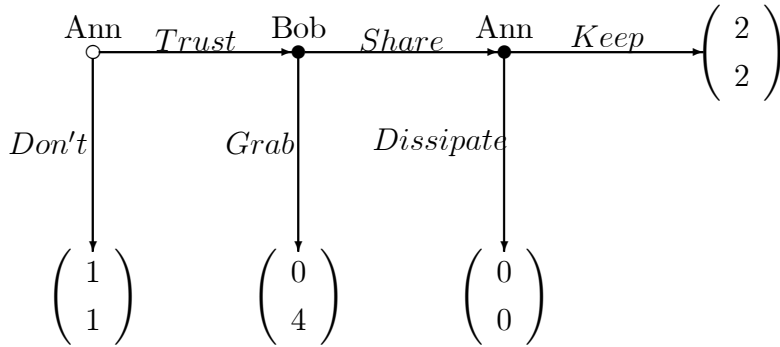
<sup>13</sup>The models can be interpreted as psychological games with asymmetric information where the utility of a player depends on the terminal beliefs of another player (cf. 6.2).

<sup>14</sup>We do not suggest that  $\Gamma_3$  is interesting only in providing a convenient alternative way to analyzing  $\Gamma_2$ ; the emotion modeled in  $\Gamma_3$  may make sense in its own right, as a primitive assumption about preferences (akin to example 3 of the introduction).

ing: how many layers of beliefs about beliefs/choices are involved) with another object involving one degree lower order. This may allow one to work with utilities of the form  $u_i : Z \times \prod_{j \neq i} (\mathbf{M}_j \times S_j) \rightarrow \mathbb{R}$ , where  $\mathbf{M}_i$  is *not* a factor of the domain. This has two methodological advantages. First, it may seem easier to work with lower order beliefs (like  $\alpha$  in  $\Gamma_3$  rather than  $\beta$  in  $\Gamma_2$ ). Second, and most importantly, one is lead to clearly distinguish between the carriers of utility (*i.e.*, elements of  $Z \times \prod_{j \neq i} (\mathbf{M}_j \times S_j)$ ) and how a player deals with uncertainty by making updated probabilistic predictions (described by elements of  $\mathbf{M}_i$ ). By contrast, when the domain of  $i$ 's utility is  $Z \times \mathbf{M}_i \times \prod_{j \neq i} (\mathbf{M}_j \times S_j)$  elements of  $\mathbf{M}_i$  end up serving both purposes.

### R3: dependence on strategies

Many forms of belief-dependent motivation require preferences to depend on overall strategies, beyond how strategies cause terminal histories. Consider  $\Gamma_4$ , a variation of  $\Gamma_1$  where Ann may 'dissipate' some payoff. The payoffs of  $\Gamma_4$  are material, not necessarily reflecting utilities.



**Figure 4.** Modified Trust Game  $\Gamma_4$  with Material Payoffs.

Recall (from the discussion of R1) the terminology that 'Ann is let down' if the material payoff she gets is less than what she expected. Now note that in  $\Gamma_4$  what she expects to get does not only depend on her beliefs about Bob, but also on how she plans to play the game. Suppose Ann plans to trust Bob and then keep the surplus if she is given the opportunity. Then her subjectively expected material payoff is  $2\alpha$ , where  $\alpha$  is the probability Ann assigns to Bob's strategy *Share if Trust*. But if she plans to trust Bob and then dissipate the surplus, her expected material payoff is zero independently of her beliefs about Bob.

Again assume that Bob suffers from guilt to the extent that he believes he lets Ann down. One way to model this is to let his utility at  $(Trust, Grab)$  be  $4 - 5\alpha$  (as in  $\Gamma_3$ ) if Ann plans to *Keep*, but 4 if she plans to *Dissipate*.<sup>15</sup>

<sup>15</sup>Note also that Ann's anticipation of feeling let down might affect her initial decision. This

This example illustrates the following: Psychological motivations often exhibit a concern for other players' intentions. Intentions depend on beliefs as well as on strategies, and the latter dependence goes beyond what is implied by how strategies induce endnodes. Therefore, the domain of our psychological utility function includes (conditional) beliefs and strategies of other players, on top of terminal histories and own beliefs.

#### **R4: non-equilibrium analysis**

**R1-R3** concern features of players' motivation one may wish to incorporate in a formal framework. The next step is to predict play. We propose a generalization of Kreps & Wilson's (1982) sequential equilibrium. We postpone illustrations until we formally introduce the concept in section 4.

While much of economic theory presumes that players coordinate on an equilibrium, it is not always clear such an assumption is justified. For one thing, people may be quite rational, and confident in others' rationality, even if they fail to coordinate. In conventional game theory, related matters have inspired work on the implications of common belief of rationality; see *e.g.* Bernheim's (1984) and Pearce's (1984) work on rationalizability. This brings us to **R4**. There is little reason to assume that equilibrium coordination is easier in psychological games than in standard games. In fact, since psychological games often seem more complicated, and since problems of equilibrium multiplicity may be enhanced, assuming equilibrium may be assuming too much *especially* in psychological games.<sup>16</sup>

Giving up the equilibrium assumption does not necessarily mean giving up on predictive power. Refer back to the psychological forward induction argument, presented for  $\Gamma_2$ . Ann and Bob perform deductive reasoning regarding one another's behavior and beliefs, and a clear-cut prediction results despite that no presumption of equilibrium is made. However, the story told was informal, and specific to  $\Gamma_2$  (or, equivalently,  $\Gamma_3$ ). It is natural to wonder about generally applicable formalizations. In section 5, we develop a framework for analyzing interactive epistemology in psychological games, without postulating equilibrium play. This is a relatively small step because our very definition of psychological game already

---

can be modeled by letting Ann's utility at (*Trust*, *Grab*) be affected by her initial beliefs and her own strategy. We pursue this point and its ramifications in section 6.

<sup>16</sup>Another reason to feel skeptical about equilibrium analysis in psychological games is the following: It is often argued that players learn to play some equilibrium because through recurrent play they come to hold correct beliefs about the opponents' actions (see, *e.g.*, Fudenberg & Levine, 1998, and references therein). This is *not enough* for psychological games; since payoffs depend on hierarchical beliefs, players would have to be able to learn others' beliefs, but unlike actions beliefs are typically not observable *ex post*.

provides the necessary ingredients. Building on an epistemic theme due to Battigalli & Siniscalchi (2002), we extend Pearce’s (1984) classical notion of (extensive form) rationalizability to psychological games. The concept captures psychological forward induction in simple games like  $\Gamma_2$  and  $\Gamma_3$ , and in more complicated games for which long chains of beliefs about beliefs are needed to get sharp predictions.

### Other modifications of GPS

**R1-R4** do not exhaust the good reasons to generalize GPS, but in the name of pedagogical clarity we only deal with **R1-R4** in sections 3-5. In section 6 we extend the perspective, to incorporate incomplete and imperfect information, chance moves, dependence of utility on own strategy, and multi-self utility.

## 3 Psychological Games

In this section we introduce notation on extensive-form games (3.1), model a universal belief space that accounts for updated beliefs (3.2), and put forth and illustrate our general definition of a psychological game (3.3).

### 3.1 Extensive forms with observable actions

We first restrict attention to finite multi-stage games with observable actions, no chance moves, and complete information. These restrictions can be removed, at the cost of additional notational complexity (see section 6). We assume players move simultaneously at every stage. This is without loss of generality, because the set of feasible actions of a player may depend on actions chosen in previous stages and may be singleton. Simultaneous moves games, perfect information games, and repeated games are special cases (cf. Osborne & Rubinstein, 1994, ch. 6). We use the following notation/terminology:

An *extensive form* with observable actions is a tuple  $\langle N, H \rangle$  where  $N = \{1, \dots, n\}$  is the *player* set, and  $H$  is the set of feasible *histories*. A history of length  $\ell$  is a sequence  $h = (a^1, \dots, a^\ell)$  where each  $a^t = (a_1^t, \dots, a_n^t)$  represents the profile of actions chosen at stage  $t$  ( $1 \leq t \leq \ell$ ). We assume history  $h$  becomes public information as soon as it occurs. We also assume  $H$  is finite. For notational convenience, we let  $H$  contain the *empty history* (of length 0), denoted  $h^0$ . The set of feasible actions for player  $i$  at history  $h$  is denoted  $A_i(h)$  and may be singleton, meaning that  $i$  is not active at  $h$ .  $A_i(h)$  is empty if and only if  $h$  is a *terminal* history.  $Z$  denotes the set of terminal histories.

For any given extensive form, we let  $S_i$  denote the set of (pure) strategies of player  $i$ . A typical strategy is denoted by  $s_i = (s_{i,h})_{h \in H \setminus Z}$ , where  $s_{i,h}$  is the action that would be selected by strategy  $s_i$  if history  $h$  occurred. Define  $S = \prod_{i \in N} S_i$  and  $S_{-i} = \prod_{j \neq i} S_j$ . The set of strategies of player  $i$  that allow history  $h$  is denoted  $S_i(h)$ . Similar notation is used for strategy profiles:  $S(h) = \prod_{i \in N} S_i(h)$  and  $S_{-i}(h) = \prod_{j \in N} S_j(h)$ . We let  $\zeta(s) \in Z$  denote the terminal history induced by strategy profile  $s = (s_i)_{i \in N}$ .

### 3.2 Conditional beliefs & infinite hierarchies of beliefs

Here we summarize the theory of hierarchies of conditional beliefs due to Battigalli & Siniscalchi (1999), which should be consulted for proofs, details, and further references. Consider a decision maker DM who is uncertain about which element in a set  $X$  is true. Assume  $X$  is a compact Polish space.<sup>17</sup> DM assigns probabilities to events  $E, F, \dots$  in the Borel sigma-algebra  $\mathcal{B}$  of  $X$  according to some (countably additive) probability measure. Let  $\Delta(X)$  denote the set of all such probability measures. As events unfold DM updates her beliefs. The actual and/or potential beliefs of DM are described by a conditional probability system (see R enyi, 1955). Let  $\mathcal{C} \subseteq \mathcal{B}$  denote the collection of potentially observable events (or conditioning events). DM holds probabilistic beliefs conditional on each event  $F \in \mathcal{C}$ .

**Definition 1** A conditional probability system (cps) on  $(X, \mathcal{B}, \mathcal{C})$  is a function  $\mu(\cdot|\cdot) : \mathcal{B} \times \mathcal{C} \rightarrow [0, 1]$  such that for all  $E \in \mathcal{B}, F, F' \in \mathcal{C}$

- (1)  $\mu(\cdot|F) \in \Delta(X)$ ,
- (2)  $\mu(F|F) = 1$ ,
- (3)  $E \subseteq F' \subseteq F$  implies  $\mu(E|F) = \mu(E|F')\mu(F'|F)$ .

We regard the set of cps' on  $(X, \mathcal{B}, \mathcal{C})$  as a subset of the topological space  $[\Delta(X)]^{\mathcal{C}}$ , where  $\Delta(X)$  is endowed with the topology of weak convergence of measures and  $[\Delta(X)]^{\mathcal{C}}$  is endowed with the product topology.

From now on DM is a player  $i$ ;  $(X, \mathcal{B}, \mathcal{C})$  is either  $X = S_{-i}$  (a finite set) or  $X = S_{-i} \times Y$  where  $Y$  is a compact Polish parameter space typically representing a set of opponents' beliefs. The Borel sigma-algebra  $\mathcal{B}$  is implicitly understood,<sup>18</sup> and conditioning events corresponds to histories, *i.e.*,  $\mathcal{C} = \{F \subseteq S_{-i} \times Y : F = S_{-i}(h) \times Y, h \in H\}$  (or  $\mathcal{C} = \{F \subseteq S_{-i} : F = S_{-i}(h), h \in H\}$  if  $X = S_{-i}$ ). The set of cps' is denoted

<sup>17</sup>A topological space  $X$  is *Polish* if it admits a compatible metric  $d$  such that  $(X, d)$  is a complete and separable metric space (see, *e.g.*, Kechris, 1995, p. 13).

<sup>18</sup> $\mathcal{B}$  obtains from the product of the discrete topology on  $S_{-i}$  and the topology of  $Y$ .

$\Delta^H(S_{-i} \times Y)$  a subset of  $[\Delta(S_{-i} \times Y)]^H$ . If conditioning event  $F$  corresponds to history  $h$ , then we abbreviate as  $\mu(\cdot|F) = \mu(\cdot|h)$ .

The following result shows that we can take for granted that  $\Delta^H(S_{-i} \times Y)$  is a compact Polish space, just like  $Y$ .<sup>19</sup> It is key in our construction of hierarchical conditional beliefs, implying that the domains of higher- and lower-order uncertainty have the same structural properties.

**Lemma 2**  *$\Delta^H(S_{-i})$  is a compact Polish space. Furthermore, if  $Y$  is a compact Polish space, also  $\Delta^H(S_{-i} \times Y)$  is a compact Polish space.*

Hierarchies of cps' are defined recursively as follows:

- $X_{-i}^0 = S_{-i}$  ( $i \in N$ ),
- $X_{-i}^k = X_{-i}^{k-1} \times \prod_{j \neq i} \Delta^H(X_{-j}^{k-1})$  ( $i \in N$ ;  $k = 1, 2, \dots$ ).

By repeated applications of Lemma 2, each  $X_{-i}^k$  is a cross-product of compact Polish spaces, hence compact Polish itself.<sup>20</sup> A cps  $\mu_i^k \in \Delta^H(X_{-i}^{k-1})$  is called  $k$ -order cps. For  $k > 1$ ,  $\mu_i^k$  is a joint cps on the opponents' strategies and  $(k-1)$ -order cps'. A *hierarchy of cps'* is a countably infinite sequence of cps'  $\boldsymbol{\mu}_i = (\mu_i^1, \mu_i^2, \dots) \in \prod_{k>0} \Delta^H(X_{-i}^{k-1})$ .  $\boldsymbol{\mu}_i$  is *coherent* if the cps' of distinct orders assign the same conditional probabilities to lower-order events:

$$\mu_i^k(\cdot|h) = \text{marg}_{X_{-i}^{k-1}} \mu_i^{k+1}(\cdot|h) \quad (k = 1, 2, \dots; h \in H).$$

It can be shown that a coherent hierarchy  $\boldsymbol{\mu}_i$  induces a cps  $\nu_i$  on the cross-product of  $S_{-i}$  with the sets of hierarchies of cps' of  $i$ 's opponents, a compact Polish space.

However,  $\nu_i$  may assign positive probability (conditional on some  $h$ ) to opponents' incoherence. To rule this out, say that a coherent hierarchy  $\boldsymbol{\mu}_i$  satisfies belief in coherency of order 1 if the induced cps  $\nu_i$  is such that each  $\nu_i(\cdot|h)$  ( $h \in H$ ) assigns probability one to the opponents' coherency;  $\boldsymbol{\mu}_i$  satisfies belief in coherency of order  $k$  if it satisfies belief in coherency of order  $k-1$  and the induced cps  $\nu_i$  is such that each  $\nu_i(\cdot|h)$  ( $h \in H$ ) assigns probability one the opponents' coherency of order  $k-1$ ;  $\boldsymbol{\mu}_i$  is *collectively coherent* if it satisfies belief in coherency of order  $k$  for each positive integer  $k$ . The set of collectively coherent hierarchies of player  $i$  is a compact Polish space, denoted by  $\mathbf{M}_i$ . We let  $M_i^k$  denote the set of  $k$ -order beliefs

<sup>19</sup>This depends on two facts: (1) the collection of conditioning events for player  $i$  (corresponding to  $H$ ) is at most countable (indeed finite), and (2) each conditioning event  $S_{-i}(h) \times Y$  (or  $S_{-i}(h)$  if  $X = S_{-i}$ ) is both closed and open.

<sup>20</sup>The cross-product of countably many compact Polish spaces is also compact Polish.

consistent with collective coherency, that is, the projection of  $\mathbf{M}_i$  on  $\Delta^H(X_{-i}^{k-1})$ , and let  $M_{-i}^k = \prod_{j \neq i} M_j^k$ ,  $\mathbf{M}_{-i} = \prod_{j \neq i} \mathbf{M}_j$ ,  $\mathbf{M} = \prod_{j \in N} \mathbf{M}_j$ .

We have now defined all the elements that form the domain of the utility functions. But is this enough for the analysis of strategic reasoning? In order to decide on the best course of action, player  $i$  may need to form (conditional) beliefs about the *infinite* hierarchies of (conditional) beliefs of other players, either because they enter his psychological payoff function or because his assessment of the behavior and finite-order beliefs of other players is derived from assumptions, such as “common belief in rationality”, involving beliefs of infinitely many orders. Does this mean that we need additional layers of beliefs? No. The following result shows that the countably infinite hierarchies of cps’ defined above are sufficient for the strategic analysis;  $\mathbf{M}_i$  is isomorphic to  $\Delta^H(S_{-i} \times \mathbf{M}_{-i})$ , so each  $\boldsymbol{\mu}_i \in \mathbf{M}_i$  corresponds to a cps on  $S_{-i} \times \mathbf{M}_{-i}$ :

**Lemma 3** *For each  $i \in N$  there is a 1-to-1 and onto continuous function*

$$f_i = (f_{i,h})_{h \in H} : \mathbf{M}_i \rightarrow \Delta^H(S_{-i} \times \mathbf{M}_{-i})$$

*whose inverse is also continuous. Furthermore, each coordinate function  $f_{i,h}$  is such that for all  $\boldsymbol{\mu}_i = (\mu_i^1, \mu_i^2, \dots) \in \mathbf{M}_i$ ,  $k \geq 1$*

$$\mu_i^k(\cdot|h) = \text{marg}_{S_{-i} \times M_{-i}^1 \times \dots \times M_{-i}^{k-1}} f_{i,h}(\boldsymbol{\mu}_i).$$

### 3.3 Psychological Games

We are now ready to state our definition of a psychological game:

**Definition 4** *A psychological game based on extensive form  $\langle N, H \rangle$  is a structure  $\Gamma = \langle N, H, (u_i)_{i \in N} \rangle$  where  $u_i : Z \times \mathbf{M} \times S_{-i} \rightarrow \mathbb{R}$  is  $i$ ’s (measurable and bounded) psychological payoff function.*

The numerical examples examined in section 2 fit this definition: in  $\Gamma_2$ ,  $u_2$  depends on  $z$  and  $\mu_2^2(\cdot|Trust)$ ;<sup>21</sup> in  $\Gamma_3$ ,  $u_2$  depends on  $z$  and 1’s initial first-order belief,  $\mu_1^1(\cdot|h^0)$ ; finally, the psychological payoff function  $u_2$  proposed to analyze  $\Gamma_4$  (a game with material payoffs) depends on  $z$ ,  $\mu_1^1(\cdot|h^0)$ , and  $s_1$ .

In all these examples, a psychological game is obtained from a *material payoff game*  $\langle N, H, (\pi_i : Z \rightarrow \mathbb{R})_{i \in N} \rangle$  according to some formula. We now illustrate a few such derivations, focusing on two-player games.

---

<sup>21</sup>  $\mu_2^2(\cdot|Trust)$  is the conditional second-order belief of player 2 used to compute the expectation  $\beta$  of the probability  $\alpha$  initially assigned by 1 to the strategy ‘Share if Trust’.

Player  $j$  is ‘let down’ if her actual material payoff  $\pi_j(z)$  is lower than the payoff she expected to get, given her initial first-order beliefs  $\mu_j^1(\cdot|h^0)$  and her strategy  $s_j$ . This disappointment can be measured by the function

$$D_j(z, \mu_j^1(\cdot|h^0), s_j) = \max \left\{ 0, \left[ \sum_{s'_i} \mu_j^1(s'_i|h^0) \pi_j(\zeta(s_j, s'_i)) - \pi_j(z) \right] \right\}.$$

A guilt motivation can be modeled as aversion to letting the other player down, which can be captured by the following payoff function:<sup>22</sup>

$$u_i(z, \boldsymbol{\mu}, s_j) = \pi_i(z) - \theta_i D_j(z, \mu_j^1(\cdot|h^0), s_j)$$

For the special case of  $\Gamma_1$ , we obtain  $\Gamma_3$  by letting  $\theta_1 = 0$  and  $\theta_2 = \frac{5}{2}$ .

Another psychological motivation we can model is the desire to avoid regret. Regret of player  $i$  at a terminal history  $z$  can be captured by the distance between the actual material payoff  $\pi_i(z)$  and the maximal expected payoff that could have been obtained ‘with the benefit of hindsight,’ *i.e.*, using the terminal beliefs conditional on  $z$ . Formally,  $i$ ’s regret equals

$$R_i(z, \boldsymbol{\mu}_i^1(\cdot|z)) = \max_{s_i} \sum_{s'_j \in S_j(z)} \mu_i^1(s'_j|z) \pi_i(\zeta(s_i, s'_j)) - \pi_i(z)$$

and we obtain the psychological payoff function

$$u_i(z, \boldsymbol{\mu}, s_j) = \pi_i(z) - \theta_i R_i(z, \boldsymbol{\mu}_i^1(\cdot|z))$$

where  $\theta_i$  is a psychological sensitivity parameter.<sup>23</sup> This shows that it may be natural to let utility depend on what players believe at the end of the game (for other examples of this kind see subsection 6.2).

---

<sup>22</sup>We opt for a formulation where  $i$ ’s psychological payoff depends directly on  $j$ ’s beliefs; cf. the discussion of R2 in section 2. Only initial beliefs enter the utility function directly, but in the strategic analysis the updated second-order beliefs of  $i$  are crucial because they determine the expected payoff  $i$  maximizes at each history. The reader may want to compare the formulation here to that of Charness & Dufwenberg (2004), which utilizes GPS’ framework and lets  $i$ ’s own initial beliefs influence  $i$ ’s utility.

<sup>23</sup>Bell (1982) and Loomes & Sugden (1982) develop theories of regret, in which a decision maker’s experienced utility depends on the post-choice revelation of a state-of-nature. Our formulation preserves that spirit, but extends it to belief-dependent motivation. This is natural in a strategic setting, where players cannot perfectly observe *ex post* the state of the world, which includes what another player *would* have chosen.

## 4 Equilibrium Analysis

Kreps & Wilson’s notion of sequential equilibrium has become a benchmark for standard games. Our goal here is to extend this concept to the class of psychological games defined in section 3. (The restriction to multi-stage game forms with complete information simplifies, but is not essential as we discuss in section 6.) We next comment on the entailed interpretation of mixed strategies and assessments (4.1), give the main definition (4.2), and consider examples (4.3).

### 4.1 Randomized strategies and consistent assessments

The equilibrium concept we develop refers to randomized choices. However, in our interpretation, we exclude explicit randomization (players tossing coins or spinning roulette wheels). Rather, we will interpret a randomized choice of a given player  $i$  as the common first-order belief of  $i$ ’s opponents about  $i$  (cf. Aumann & Brandenburger, 1995). This is the analog of the following characterization of a Nash equilibrium in a standard simultaneous moves game: a profile  $(\sigma_1, \dots, \sigma_n) \in \Delta(A_1) \times \dots \times \Delta(A_n)$  is an equilibrium if for each player  $i$  each action in the support of  $\sigma_i$  is a best response to  $\sigma_{-i}$ .

We focus on behavior strategies  $\sigma_i = (\sigma_i(\cdot|h))_{h \in H \setminus Z} \in \prod_{h \in H \setminus Z} \Delta(A_i(h))$ , interpreting  $\sigma_i$  as an array of common conditional first-order beliefs held by  $i$ ’s opponents. This interpretation is part of the notion of ‘consistency’ of profiles of strategies and hierarchical beliefs defined below.

Kreps & Wilson argue that an appropriate definition of equilibrium in extensive form games must refer to ‘assessments: profiles of (behavior) strategies *and* conditional (first-order) beliefs. They define sequential equilibrium in two steps: first a ‘consistency’ condition for assessments, and then sequential equilibrium is a consistent assessment that satisfies sequential rationality. It turns out that the consistency condition captures the assumptions that (a) each player regards his opponents’ strategies as stochastically independent, and (b) any two players have the same (prior and conditional) beliefs about any third player (cf. Fudenberg & Tirole 1991b, Battigalli 1996, Kohlberg & Reny 1997). We follow a similar approach, adding a third requirement concerning the higher-order beliefs that need to be specified in psychological games.

In our setup, an *assessment* is a profile  $(\sigma, \boldsymbol{\mu}) = (\sigma_i, \boldsymbol{\mu}_i)_{i \in N}$  of (behavior) strategies and hierarchies of conditional beliefs. Before defining consistency of an assessment, we need to define more precisely what we mean by ‘stochastic independence’. For this, we need to explain that a *marginal cps* on the strategies

of  $j$  is a cps on  $(S_j, \mathcal{B}_j, \mathcal{C}_j)$ , where  $\mathcal{B}_j$  is the power set of  $S_j$  and  $\mathcal{C}_j = \{S_j(h), h \in H\}$ . The set of marginal cps' is denoted by  $\Delta^H(S_j)$ . The following definition takes advantage of the simple information structure we are assuming, *i.e.* perfect observability of past actions, and allows us to characterize stochastic independence for cps's in terms of 'marginal' cps's.

**Definition 5** *A first-order cps  $\mu_i \in \Delta^H(S_{-i})$  satisfies stochastic independence, if there exists a profile of marginal cps's  $(\mu_{ij})_{j \neq i} \in \prod_{j \neq i} \Delta^H(S_j)$  s.t.  $\mu_i(s_{-i}|h) = \prod_{j \neq i} \mu_{ij}(s_j|h)$  for all  $h \in H$ ,  $s_{-i} \in S_{-i}(h)$ . We let  $\Delta_I^H(S_{-i})$  denote the set of first-order cps's of player  $i$  that satisfy stochastic independence.*

Note that for each  $\mu_i \in \Delta_I^H(S_{-i})$  we can derive a behavior strategy profile  $(\sigma_j)_{j \neq i}$  as follows: let  $S_j(h, a_j) = \{s_j \in S_j(h) : s_{j,h} = a_j\}$  denote the set of strategies of player  $j$  that allow history  $h$  and select action  $a_j$  at  $h$ , then

$$\forall j \neq i, \forall h \in H, \forall a_j \in A_j(h), \sigma_j(a_j|h) = \mu_{ij}(S_j(h, a_j)|h). \quad (1)$$

We are now ready for the main definition of this section:

**Definition 6**  $\boldsymbol{\mu} = (\boldsymbol{\mu}_i)_{i \in N} \in \mathbf{M}$  is consistent if

(a) the first-order cps of each player satisfies stochastic independence:

$$\forall i \in N, \mu_i^1 \in \Delta_I^H(S_{-i}),$$

(b) the marginal first-order beliefs of two players about a third coincide:

$$\forall i, \forall j \in N, \forall k \in N \setminus \{i, j\}, \forall h \in H, \text{marg}_{S_k} \mu_i^1(\cdot|h) = \text{marg}_{S_k} \mu_j^1(\cdot|h),$$

(c) each player's higher-order beliefs in  $\boldsymbol{\mu}$  assign probability one to the lower-order beliefs in  $\boldsymbol{\mu}$  itself:

$$\forall i \in N, \forall k > 1, \forall h \in H, \mu_i^k(\cdot|h) = \mu_i^{k-1}(\cdot|h) \times \delta_{\mu_{-i}^{k-1}}$$

where  $\delta_x$  is the measure that assigns probability one to the singleton  $\{x\}$ . An assessment  $(\sigma, \boldsymbol{\mu})$  is consistent if  $\boldsymbol{\mu}$  is consistent and  $\sigma$  is derived from first-order beliefs  $(\mu_i^1)_{i \in N}$  as in eq. (1).

The justification of the (strong) conditions (b) and (c) comes from the classical interpretation of equilibrium beliefs as the end-product of a transparent reasoning process by intelligent players. Therefore any two players must share the same

first-order conditional beliefs about any other player, and every player comes to a correct conclusion about the (hierarchical) beliefs of his opponents because he is able to replicate their reasoning.<sup>24</sup> Condition (c) is analogous to a condition used by GPS to define psychological Nash equilibrium, requiring that players hold common, correct beliefs about each others' beliefs. This is equivalent to the requirement that, for each player  $i$  and each history  $h$ , the conditional belief on  $S_{-i} \times \mathbf{M}_{-i}$  induced by hierarchy  $\boldsymbol{\mu}_i$  assigns probability one to  $\boldsymbol{\mu}_{-i}$ .<sup>25</sup>

## 4.2 Sequential Equilibrium Assessments

We take the point of view of an 'agent'  $(i, h)$  of player  $i$ , in charge of the move at history  $h$ , who seeks to maximize  $i$ 's conditional expected utility given the consistent belief profile  $\boldsymbol{\mu}$ . The expected utility of  $i$  conditional on history  $h$  and action  $a_i \in A_i(h)$  given  $\boldsymbol{\mu}$  can be expressed as

$$E_{\boldsymbol{\mu}}[u_i|h, a_i] = \sum_{s_{-i} \in S_{-i}(h)} \mu_i^1(s_{-i}|h) \sum_{s_i \in S_i(h, a_i)} \mu_{ji}^1(s_i|(h, a_i, s_{-i, h})) u_i(\zeta(s), \boldsymbol{\mu}, s_{-i}) \quad (2)$$

where  $s_{-i, h}$  is the action profile chosen by  $i$ 's opponents at  $h$  according to  $s_{-i}$  and  $\mu_{ji}^1$  is the first-order cps about  $i$  of an arbitrary opponent  $j$ . This specification presumes that  $(i, h)$  assesses the probabilities of actions by other agents of player  $i$  in the same way as each player  $j \neq i$ ; that explains how  $\mu_{ji}^1(s_i|(h, a_i, s_{-i, h}))$  shows up in the right-hand-side of the expression.

The expected utility formula (2) is quite different from those used in the literature on standard games. This is because of the possibility that psychological payoffs are directly affected by strategies. If this is ruled out,  $E_{\boldsymbol{\mu}}[u_i|h, a_i]$  can be expressed in a more familiar form:

**Remark 7** *Suppose that psychological payoff functions depend only on terminal*

---

<sup>24</sup>Condition (c) implies that although players update their beliefs about opponents' strategies they never change their beliefs about what opponents would believe conditional on each history. Of course, by observing the actual play-path each player infers the current actual beliefs of his opponents, but interesting forms of learning about others' beliefs are ruled out. For example, (c) implies that no player would ever change his mind about his opponents' initial beliefs. Without defending this assumption, we argue that it is in the spirit of the standard definition of sequential equilibrium. After a hypothetical deviation by  $i$ , this player is assumed to play a continuation strategy maximizing his expected payoff against the same (equilibrium) beliefs that were ascribed to him before the deviation, even if the deviation is irrational under such beliefs (cf. Reny, 1992).

<sup>25</sup>That is, (3) holds iff  $\forall i \in N, \forall h \in H, f_{i, h}(\boldsymbol{\mu}_i)(S_{-i} \times \{\boldsymbol{\mu}_{-i}\}) = 1$ .

histories and beliefs. Then for any consistent assessment  $(\sigma, \boldsymbol{\mu})$

$$E_{\boldsymbol{\mu}}[u_i|h, a_i] = \sum_z \Pr_{\sigma}[z|h, a_i] u_i(z, \boldsymbol{\mu})$$

where  $\Pr_{\sigma}[z|h, a_i]$  is the probability of terminal history  $z$  conditional on  $(h, a_i)$  determined by behavioral profile  $\sigma$ .

We now move to the section's main definition. A consistent assessment is a sequential equilibrium if it satisfies a sequential rationality condition:

**Definition 8** *Assessment  $(\sigma, \boldsymbol{\mu}) = (\sigma_i, \boldsymbol{\mu}_i)_{i \in N}$  is a sequential equilibrium (SE) if it is consistent and for all  $i \in N$ ,  $h \in H \setminus Z$ ,*

$$\text{Supp}(\sigma_i(\cdot|h)) \subseteq \arg \max_{a_i \in A_i(h)} E_{\boldsymbol{\mu}}[u_i|h, a_i]. \quad (3)$$

The sequential rationality condition (3) only requires immunity to ‘one-shot’ deviations. By application of the ‘one-shot-deviation principle’, this is equivalent to immunity to deviations to arbitrary continuation strategies.<sup>26</sup> The following proposition stresses this point and the ‘belief interpretation’ of randomization in our definition of SE (the proof is available on request).

Fix a hierarchy of cps’  $\boldsymbol{\mu}_i$  a (non terminal) history  $h$  and a strategy  $s_i$  consistent with  $h$ . The expectation of  $u_i$  conditional on  $h$ , given  $s_i$  and  $\boldsymbol{\mu}_i$  is

$$E_{s_i, \boldsymbol{\mu}_i}[u_i|h] = \int_{S_{-i} \times \mathbf{M}_{-i}} u_i(\zeta(s_i, s_{-i}), \boldsymbol{\mu}_i, \boldsymbol{\mu}_{-i}, s_{-i}) f_{i,h}(\boldsymbol{\mu}_i)(ds_{-i}, d\boldsymbol{\mu}_{-i}). \quad (4)$$

**Proposition 9**  *$\boldsymbol{\mu} = (\boldsymbol{\mu}_i)_{i \in N}$  is part of a sequential equilibrium assessment if and only if  $\boldsymbol{\mu}$  is consistent and for all  $i \in N$ ,  $h \in H \setminus Z$ ,  $j \in N \setminus \{i\}$ ,*

$$\text{Supp}\boldsymbol{\mu}_{j_i}^1(\cdot|h) \subseteq \arg \max_{s_i \in S_i(h)} E_{s_i, \boldsymbol{\mu}_i}[u_i|h].$$

The main result of this section is an existence theorem:

**Theorem 10** *If the psychological payoff functions are continuous, there exists at least one sequential equilibrium assessment.*

---

<sup>26</sup>The ‘one-shot-deviation principle’ is essentially a dynamic programming result. It holds for finite (standard) games, and more generally for finite-horizon games, and infinite-horizon games where payoffs are ‘continuous at infinity’. See, e.g., Fudenberg & Tirole (1991a), pp 108-110. Subsection 6.4 shows that when the definition of psychological game is extended to allow  $u_i$  to depend on  $s_i$  the one-shot-deviation principle does not apply.

The proof is rather straightforward but somewhat tedious. We provide a sketch here (but details are available on request). Consider  $\varepsilon$ -perturbed games where there is positive minimal probability of choosing any action at any history, *i.e.*  $\varepsilon = (\varepsilon_{i,h}(a_i, h))_{a_i \in A_i(h), i \in N, h \in H}$  is a strictly positive vector such that  $\sum_{a_i \in A_i(h)} \varepsilon(a_i, h) < 1$  for each history  $h$  (cf. Selten, 1975). For each strictly positive behavior strategy profile, there exists a corresponding profile of hierarchies of cps'  $\boldsymbol{\mu} = \beta(\sigma)$  such that  $(\sigma, \beta(\sigma))$  is consistent.<sup>27</sup> For any  $\varepsilon$ -perturbed game we define an (agent-form, psychological)  $\varepsilon$ -equilibrium as an  $\varepsilon$ -constrained behavior strategy profile  $\sigma_\varepsilon$  such that for each history  $h$  and each player  $i$ , a pure action  $a_i$  that does not maximize the expectation of  $u_i$  (given  $h$  and  $\beta(\sigma_\varepsilon)$ ) is assigned the minimal probability  $\varepsilon(a_i, h)$ . It can be shown by standard compactness-continuity arguments that each  $\varepsilon$ -perturbed game has an  $\varepsilon$ -equilibrium (cf. the existence proof for psychological Nash equilibria in GPS). Fix a sequence  $\varepsilon^k \rightarrow 0$  and a corresponding sequence  $\sigma^k$  of  $\varepsilon^k$ -equilibrium assessments. By compactness,  $\sigma^k$  has an accumulation point  $\sigma^*$ . By upper-hemicontinuity of the local best response correspondences, for each  $(i, h)$ ,  $\sigma_i^*(\cdot|h)$  assigns positive probability only to actions that are best responses to  $(\sigma^*, \beta(\sigma^*))$  at  $h$ . Therefore  $(\sigma^*, \beta(\sigma^*))$  is a sequential equilibrium assessment.

We next show that the SE concept generalizes subgame perfect equilibrium for standard games with observable actions (recall: sequential and subgame perfect equilibrium coincide in games with observable actions). This is a corollary of a more general result for games where psychological payoffs depend only on terminal nodes and beliefs:  $u_i : Z \times \mathbf{M} \rightarrow \mathbb{R}$ . For any such game  $\Gamma = \langle N, H, (u_i)_{i \in N} \rangle$  and any profile of hierarchies of cps'  $\boldsymbol{\mu} = (\boldsymbol{\mu}_i)_{i \in N}$ , we can obtain a *standard* game  $\Gamma^\boldsymbol{\mu} = \langle N, H, (v_i^\boldsymbol{\mu})_{i \in N} \rangle$  with payoff functions  $v_i^\boldsymbol{\mu}(z) = u_i(z, \boldsymbol{\mu})$ .

**Proposition 11** *Suppose that psychological payoff functions have the form  $u_i : Z \times \mathbf{M} \rightarrow \mathbb{R}$ . Then an assessment  $(\sigma, \boldsymbol{\mu})$  is a sequential equilibrium if and only if it is consistent and  $\sigma$  is a subgame perfect (hence sequential) equilibrium of the standard game  $\Gamma^\boldsymbol{\mu}$ .*

**Proof.** First recall that when payoffs do not directly depend on strategies, the conditional expected payoffs determined by a consistent assessment  $(\sigma, \boldsymbol{\mu})$  can be expressed as  $E_\boldsymbol{\mu}[u_i|h, a_i] = \sum_z \Pr_\sigma[z|h, a_i] u_i(z, \boldsymbol{\mu})$  (see remark 7). Let  $(\sigma, \boldsymbol{\mu})$  be a SE. By definition  $(\sigma, \boldsymbol{\mu})$  is consistent. Since  $\text{supp}(\sigma_i(\cdot|h)) \subseteq \arg \max_{a_i \in A_i(h)} E_\boldsymbol{\mu}[u_i|h, a_i]$

<sup>27</sup>By Kuhn's transformation, a strictly positive behavior strategy profile  $\sigma$  corresponds to a product measure  $\mu_1 \times \dots \times \mu_n$  on  $S_1 \times \dots \times S_n$ ; each marginal measure  $\mu_{-i}$  on  $S_{-i}$  yields a first-order cps for  $i$  satisfying stochastic independence, and by construction the first-order cps's of different players 'agree'; a corresponding profile of hierarchies is obtained assuming that there is 'common knowledge' of beliefs, as in condition (c) of Definition 6.

for all  $i$  and  $h \in H \setminus Z$ , no player can profit from pure or randomized one-shot-deviations from  $\sigma$ . Since  $\Gamma^\mu$  is finite, the one-shot-deviation principle applies, implying  $\sigma$  is subgame perfect in standard game  $\Gamma^\mu$ . Now suppose  $(\sigma, \mu)$  is consistent; if  $\sigma$  is also a subgame perfect equilibrium of  $\Gamma^\mu$  then the sequential rationality condition (3) of Definition 8 is satisfied, so  $(\sigma, \mu)$  is a SE. ■

### 4.3 Examples

We illustrate the SE concept with three examples, first a simultaneous move game illustrating how we can reproduce the essence of a leading example of GPS, then two versions of the Trust Game connecting back to some of the key notions previously highlighted in section 2.

*Equilibrium beliefs in the Bravery Game.*

The Bravery Game is a numerical example used in GPS (p. 66) to show that a psychological game may have multiple, isolated mixed strategy equilibria even if there is only one active player, which is impossible in standard games. We consider a modified version to illustrate, in a very simple case, our definition of equilibrium in beliefs. Let  $A_1 = \{Wait\}$ ,  $A_2 = \{bold, timid\}$ . Player 1 (Ann) is inactive so we can ignore her payoff function, but her beliefs matter. Player 2 (Bob) is concerned about what Ann thinks about him. Acting boldly is dangerous, but worthwhile if Ann expects Bob to act boldly. GPS model the situation with a payoff function of the form  $\bar{u}_2 : A \times \overline{\mathbf{M}}_2 \rightarrow \mathbb{R}$ . Specifically, let  $\alpha := \mu_1^1(bold|h^0)$  denote the first-order belief of Ann about Bob (a random variable from Bob's point of view), and let  $\beta := \int \alpha \mu_2^2(d\mu_1^1)$  denote (a feature of) the second-order beliefs of Bob. The payoff function considered by GPS is

$$\bar{u}_2(a_2, \bar{\mu}_2) = \begin{cases} 2 - \beta, & \text{if } a_2 = bold \\ 3(1 - \beta), & \text{if } a_2 = timid \end{cases}$$

We modify  $\bar{u}_2$ , considering instead  $u_2 : A \times \mathbf{M} \rightarrow \mathbb{R}$  defined by

$$u_2(a_2, \mu) = \begin{cases} 2 - \alpha, & \text{if } a_2 = bold \\ 3(1 - \alpha), & \text{if } a_2 = timid \end{cases}$$

Clearly, the expectation of  $u_2$  given  $a_2$  and Bob's second-order belief  $\beta$  is  $\bar{u}_2$ . There are three equilibria, with  $\beta = \alpha = 1$ ,  $\beta = \alpha = 0$  and  $\beta = \alpha = \frac{1}{2}$ .<sup>28</sup>

---

<sup>28</sup>These are essentially the same equilibria as those obtained by GPS. But they allow for explicit randomization; thus the first-order beliefs of Player 2 are degenerate on the equilibrium (mixed)

### Trust Game with Guilt Aversion

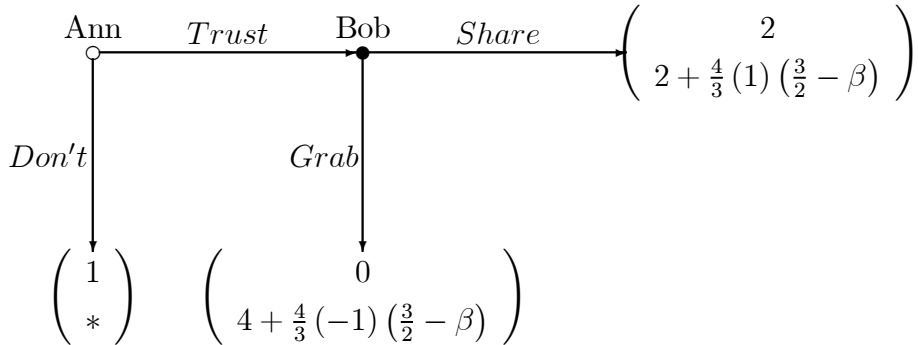
Consider  $\Gamma_3$  (or equivalently  $\Gamma_2$ ). Recall that  $\alpha$  (a function of  $\mu_1^1$ ) is the probability Ann assigns to strategy ‘Share if Trust’ at the beginning of the game, and  $\beta = \int \alpha \mu_2^2(d\mu_1^1 | Trust)$  is the relevant feature of the conditional second-order beliefs of Bob. We let  $\tau = \mu_2^1(Trust|h^0)$  denote Bob’s initial first-order belief. An assessment is summarized by  $(\tau, \alpha, \beta)$ , where  $(\tau, \alpha)$  corresponds to a behavior strategy profile. The indifference condition for Bob is  $\beta = \frac{2}{5}$ , the indifference condition for Ann is  $\alpha = \frac{1}{2}$ ; consistency yields  $\alpha = \beta$ . The game has three SEs:  $\tau = \alpha = \beta = 1$  (trust),  $\tau = \alpha = \beta = 0$  (no trust), and  $\tau = 0, \alpha = \beta = \frac{2}{5}$  (insufficient trust). Note that only the first equilibrium is consistent with forward induction reasoning (as described in section 2, and further elaborated on in subsection 5.1 below).

### Trust Game with Reciprocity

Our framework is adequate for modeling reciprocity in extensive games. To support this claim, we show how the essence of Dufwenberg & Kirchsteiger’s theory can be captured in  $\Gamma_1$ : Let  $\alpha, \beta$  and  $\tau$  be defined as in the previous example. The key tenets of the theory concern player  $i$ ’s kindness to player  $j$  ( $K_{ij}$ ), and  $i$ ’s belief in  $j$ ’s kindness to  $i$  ( $\hat{K}_{iji}$ ). At each history, player  $i$  maximizes utility defined by the sum of material payoffs (as in  $\Gamma_1$ ) and reciprocity payoffs equal to  $\theta_i \times K_{ij} \times \hat{K}_{iji}$ , where  $\theta_i$  is a constant measuring  $i$ ’s sensitivity to reciprocity. Assume that Ann’s and Bob’s sensitivities are  $\theta_1 = 0$  and  $\theta_2 = \frac{4}{3}$ . One can show that  $K_{ij}$  and  $\hat{K}_{iji}$  can be reproduced in our framework and notation; in particular we need the following:

- Bob’s kindness following *Trust* =  $-1$  for *Grab* and  $= 1$  for *Share*,
- Bob’s belief in Ann’s kindness following *Trust* =  $\frac{3}{2} - \beta$ .

$\Gamma_5$  displays the relevant utilities as conceived by the players when they move (Bob is not active at  $h^0$ , so we put no utility for him following *Don’t*):<sup>29</sup>



**Figure 5.** Trust Game  $\Gamma_5$  with Reciprocity Payoffs.

strategy of Player 1, and higher-order beliefs of each player are degenerate on the equilibrium lower-order beliefs of the other player.

<sup>29</sup>As with  $\Gamma_2$  vs.  $\Gamma_3$ , we can replace Bob’s conditional second-order belief  $\beta$  with Ann’s initial first-order belief  $\alpha$  in Bob’s payoffs, and get analogous conclusions.

Applying Definition 8,  $\Gamma_5$  has a unique SE with  $\tau = 1$ ,  $\alpha = \beta = \frac{3}{4}$ . No ‘pure’ SE exists,<sup>30</sup> just like in Dufwenberg & Kirchsteiger’s theory (cf. 6.4 below).

## 5 Interactive Epistemology

We argued in section 2 that alternatives to equilibrium analysis are worth exploring for psychological games. The definition of  $\mathbf{M}_i$  provides us with all the ingredients to analyze strategic reasoning by means of interactive epistemology assumptions, *i.e.*, assumptions about players’ rationality and what they believe about each other at any point of the game. We show how to express such assumptions in the language of events and belief operators (5.1), and then analyze a notion of extensive form rationalizability (5.2).

### 5.1 States of the world, events, and belief operators

A state of the world specifies, for each player  $i$  and history  $h$ , what  $i$  would do and believe if  $h$  were reached. Note the subjunctive conditional: game-theoretic analysis does not only concern the actual path of actions and beliefs, but also considers how players *would* react (in terms of choices and beliefs) to histories that do not actually occur at the true state. The state of a player is therefore given by his strategy and his hierarchy of cps’,  $(s_i, \boldsymbol{\mu}_i)$ . The set of states for player  $i$  is denoted by  $\Omega_i = S_i \times \mathbf{M}_i$ , and the set of *states of the world* is  $\Omega = \prod_{i=1}^n \Omega_i$ . We let  $\Omega_{-i} = \prod_{j \neq i} \Omega_j$  denote the set of possible states of  $i$ ’s opponents. With a slight abuse of notation we often write  $\Omega = \Omega_i \times \Omega_{-i}$  with typical element  $\omega = (\omega_i, \omega_{-i})$ .

An *event* is a (Borel) subset  $E \subseteq \Omega$ ; its complement is denoted  $\neg E = \Omega \setminus E$ . An event about  $i$  is any set of states  $E = E_i \times \Omega_{-i}$ , where  $E_i$  is a Borel subset of  $\Omega_i$ . We let  $\mathcal{E}_i$  denote the family of events about  $i$ . Events about the opponents of  $i$  are similarly defined; the collection of such events is denoted  $\mathcal{E}_{-i}$ .

We often use brackets to denote specific events. In particular, for any function  $\mathbf{x} : \Omega \rightarrow X$  and value  $x^* \in X$ , we use the notation  $[\mathbf{x} = x^*] := \{\omega : \mathbf{x}(\omega) = x^*\}$ . When  $\mathbf{x}$  is understood, we simply write  $[x^*]$ . For example,  $[s_i^*] = \{(s_i, \boldsymbol{\mu}_i, \omega_{-i}) : s_i = s_i^*\} \in \mathcal{E}_i$  is the event “ $i$  plays  $s_i^*$ ”, where it is understood that  $\mathbf{x}$  is the projection function on  $S_i$ , that is  $\mathbf{x}(s_i, \boldsymbol{\mu}_i, \omega_{-i}) = s_i$ . Similarly,  $[h] = \prod_{i \in N} S_i(h) \times \mathbf{M}_i$  is the event that history  $h$  occurs.

---

<sup>30</sup>In any SE we have  $\alpha = \beta$ . With  $\theta_2 = \frac{4}{3}$ , the indifference condition for Bob yields  $\beta = \frac{3}{4}$ . If  $\alpha = \beta < \frac{3}{4}$  then  $\hat{K}_{212}$  shoots up, so Bob prefers *Share* to *Grab*, which in SE would imply  $\alpha = \beta = 1$ , ... a contradiction. If  $\alpha = \beta > \frac{3}{4}$  then  $\hat{K}_{212}$  goes down, so Bob prefers *Grab* to *Share*, implying  $\alpha = \beta = 0$ , ... another contradiction.

We follow both GPS and Battigalli & Siniscalchi (2002) in disregarding players' beliefs about themselves. At state  $\omega = (s_i, \boldsymbol{\mu}_i, \omega_{-i})$ , player  $i$  would believe event  $E = \Omega_i \times E_{-i} \in \mathcal{E}_{-i}$  conditional on history  $h$  with probability  $f_{i,h}(\boldsymbol{\mu}_i)(E_{-i})$  (cf. subsection 3.2). Thus  $\{(s_i, \boldsymbol{\mu}_i, \omega_{-i}) : f_{i,h}(\boldsymbol{\mu}_i)(E_{-i}) = 1\}$  is the event “player  $i$  would believe  $E$  conditional on  $h$ ”.  $E$  itself may be an event concerning the beliefs of  $i$ 's opponents.

We use belief operators to represent events about interactive beliefs: a *belief operator* for player  $i$  is a mapping with domain  $\mathcal{E}_{-i}$  and range  $\mathcal{E}_i$ . For any given history  $h \in H$ , the  *$h$ -conditional belief operator* for  $i$  is defined as follows:

$$\forall E = \Omega_i \times E_{-i} \in \mathcal{E}_{-i}, B_{i,h}(E) = \{(s_i, \boldsymbol{\mu}_i, \omega_{-i}) : f_{i,h}(\boldsymbol{\mu}_i)(E_{-i}) = 1\}.$$

$h$  may be counterfactual at  $\omega$ , because strategies played at  $\omega$  may not induce  $h$ ; in this case “ $i$  would believe  $E$  conditional on  $h$ ” is a counterfactual statement about  $i$ 's beliefs at  $\omega$ . Clearly,  $B_{i,h}(E) \in \mathcal{E}_i$ .<sup>31</sup>  $B_{i,h}(\cdot)$  satisfies monotonicity [ $E \subseteq F$  implies  $B_{i,h}(E) \subseteq B_{i,h}(F)$ ] and conjunctiveness [ $B_{i,h}(E \cap F) = B_{i,h}(E) \cap B_{i,h}(F)$ ]. Furthermore  $B_{i,h}(E) = B_{i,h}(E \cap [h])$  because  $i$  always believes what he observes.

The basic event we are interested in is players' rationality. We take the point of view that the basic notion of rationality in extensive form games refers to plans-of-action rather than strategies; a rational player does not have to plan in advance what he would do if he deviated from his own plan. We say that player  $i$  is *rational* at state  $(s_i, \boldsymbol{\mu}_i, \omega_{-i})$  iff  $s_i$  maximizes  $i$ 's conditional expected utility  $E_{s_i, \boldsymbol{\mu}_i}[u_i|h]$  (defined in (4)) conditional on each history  $h$  allowed by  $s_i$ . More formally, let  $H_i(s_i^*) = \{h \in H \setminus Z : s_i^* \in S_i(h)\}$  denote the set of non-terminal histories allowed by  $s_i^*$ ; we require  $s_i \in r(\boldsymbol{\mu}_i)$  where

$$r_i(\boldsymbol{\mu}_i) = \left\{ s_i^* : \forall h \in H(s_i^*), s_i^* \in \arg \max_{s_i \in S_i(h)} E_{s_i, \boldsymbol{\mu}_i}[u_i|h] \right\} \quad (5)$$

The event “player  $i$  is rational” is  $R_i = \{(s_i, \boldsymbol{\mu}_i, \omega_{-i}) : s_i \in r_i(\boldsymbol{\mu}_i)\}$ . It can be shown that  $r_i(\boldsymbol{\mu}_i)$  is obtainable via a backward induction algorithm and that  $R_i$  is a well-defined nonempty event (cf. proof of Lemma 15 in the appendix).

To illustrate how these concepts can be used, we re-examine two psychological versions of the Trust Game. As regards notation, we have to distinguish the event “Bob shares”, which in the extensive form implies that “Ann trusts Bob,” from the event “Bob would share if Ann trusted Bob” which is a subjunctive conditional,

<sup>31</sup>For any Borel set  $\Omega_i \times E_{-i}$ ,  $B_{i,h}(\Omega_i \times E_{-i})$  is also a Borel set because the  $h$ -coordinate belief function  $f_{i,h}$  is continuous (see Lemma 3).

logically independent on whether Ann trusts Bob or not. Similar considerations hold for the action *Grab*. We use bold letters to denote subjunctive conditionals (which in this case correspond to strategies of Bob), as in [**Share**] and [**Grab**].

Consider the Trust Game with guilt aversion  $\Gamma_3$ . The game can be solved by forward induction reasoning: it is rational for Ann to trust Bob only if she assigns at least 50% probability to strategy **Share**, *i.e.* only if  $\alpha \geq \frac{1}{2}$ , where  $\alpha : \mathbf{M}_1 \rightarrow \mathbb{R}$  is the random variable defined by  $\alpha(\boldsymbol{\mu}_1) = \mu_1^1(\mathbf{Share}|h^0)$ .<sup>32</sup> If Bob believes in Ann's rationality when he has to move (even if he is 'surprised'), he infers from Ann's action *Trust* that  $\alpha \geq \frac{1}{2}$ . Therefore  $\beta \geq \frac{1}{2}$ , where  $\beta : \mathbf{M}_2 \rightarrow \mathbb{R}$  is the random variable defined by  $\beta(\boldsymbol{\mu}_2) = \int \alpha(\boldsymbol{\mu}_1) f_{2,Trust}(\boldsymbol{\mu}_2)(d\boldsymbol{\mu}_1)$ . His rational response is to share. If Ann anticipates Bob's reasoning she trusts him.

The formal counterpart of this argument is as follows (the events listed are nonempty; we rely on the monotonicity of the belief operators):

$$\begin{aligned} R_1 &= \left\{ (s_1, \boldsymbol{\mu}_1, \omega_2) : \alpha(\boldsymbol{\mu}_1) > \frac{1}{2} \Rightarrow s_1 = \textit{Trust}, \alpha(\boldsymbol{\mu}_1) < \frac{1}{2} \Rightarrow s_1 = \textit{Don't} \right\} \\ R_2 &= \left\{ (\omega_1, s_2, \boldsymbol{\mu}_2) : \beta(\boldsymbol{\mu}_2) > \frac{2}{5} \Rightarrow s_2 = \mathbf{Share}, \beta(\boldsymbol{\mu}_2) < \frac{2}{5} \Rightarrow s_2 = \mathbf{Grab} \right\}, \end{aligned}$$

$$R_1 \cap [\textit{Trust}] \subseteq \left[ \alpha \geq \frac{1}{2} \right],$$

$$B_{2,Trust}(R_1) = B_{2,Trust}(R_1 \cap [\textit{Trust}]) \subseteq B_{2,Trust}\left(\left[\alpha \geq \frac{1}{2}\right]\right) \subseteq \left[\beta \geq \frac{1}{2}\right],$$

$$R_2 \cap B_{2,Trust}(R_1) \subseteq R_2 \cap \left[\beta \geq \frac{1}{2}\right] \subseteq [\mathbf{Share}],$$

$$R_1 \cap B_{1,h^0}(R_2 \cap B_{2,Trust}(R_1)) \subseteq R_1 \cap [\alpha = 1] \subseteq [\textit{Trust}].$$

Now consider the Trust Game with Reciprocity  $\Gamma_5$  (or the equivalent version with  $\beta$  replaced by  $\alpha$ ). Without an equilibrium supposition, one is at loss for predictive power: if  $\theta_2 = \frac{4}{3}$ , Bob's best response depends on whether  $\beta$  is below or above  $\left(\frac{3}{2} - \frac{1}{\theta_2}\right) = \frac{3}{4}$ . This cannot be resolved by forward induction reasoning, which yields (as explained above)  $\beta \geq \frac{1}{2}$ .

However, if one uses other values of  $\theta_2$  one can draw clear conclusions merely using backward induction: if  $\theta_2 < \frac{2}{3}$ , Bob's best response (given *Trust*) is *Grab* independently of  $\beta$ , thus  $R_2 \subseteq [\mathbf{Grab}]$  and  $R_1 \cap B_{1,h^0}(R_2) \subseteq [\textit{Don't}]$ ; on the other hand, if  $\theta_2 > 2$ ,  $R_2 \subseteq [\mathbf{Share}]$  and  $R_1 \cap B_{1,h^0}(R_2) \subseteq [\textit{Trust}]$ . Furthermore, a

---

<sup>32</sup>In some formulas, we have to make explicit the dependence of random variable  $\alpha$  on the state of the world. The same holds for random variable  $\beta$ .

subtle issue arises when  $\frac{2}{3} < \theta_2 < 1$ : backward induction cannot pin down Bob’s best response, which is *Grab* if  $\beta \geq \left(\frac{3}{2} - \frac{1}{\theta_2}\right)$ , while a forward induction yields  $\beta \geq \frac{1}{2}$ . This puts an *upper bound* on how kind Bob believes Ann is,<sup>33</sup> and with  $\frac{2}{3} < \theta_2 < 1$  the best response is *Grab* (formally,  $R_2 \cap B_{2,Trust}(R_1) \subseteq [\mathbf{Grab}]$ ,  $B_{1,h^0}(R_2 \cap B_{1,Trust}(R_1)) \subseteq [\alpha = 0]$  and  $R_1 \cap B_{1,h^0}(R_2 \cap B_{1,Trust}(R_1)) \subseteq [Don't]$ ).

One can show that the SE prediction implies  $0 < \alpha = \beta = \left(\frac{3}{2} - \frac{1}{\theta_2}\right) < \frac{1}{2}$ ,  $\tau = 0$ . Thus, SE and forward induction reasoning yield the same path, but very different predictions about how Bob would revise his beliefs off that path.<sup>34</sup>

## 5.2 Rationalizability

The basic rationalizability concept for standard games is equivalent to iterated strict dominance and is motivated by the assumption that players are rational and there is common belief in rationality. Several modifications of rationalizability have been proposed, to handle sequential rationality and to reflect alternative epistemological assumptions.<sup>35</sup> In psychological games payoffs are affected by hierarchical beliefs, so rationalizability has to be defined as a property of a whole state of the world rather than of strategies. One could, *e.g.*, stipulate that a state  $\omega = (s_i, \mu_i)_{i \in N}$  is rationalizable if at  $\omega$  players are rational and there is common belief in rationality at the beginning of the game.<sup>36</sup>

One could go on to examine an array of modifications. However, that is not our goal here. Rather, we wish to indicate that the class of psychological games we have defined is amenable to interactive epistemology analysis in principle, and to illustrate the potential cutting power of such an approach. Specifically, we provide the tools to perform a forward-induction analysis of general psychological games. Following Battigalli & Siniscalchi (2002), we first define a ‘*strong belief* operator’  $SB_i$  as follows:  $SB_i(\emptyset) = \emptyset$  and

$$\forall E \in \mathcal{E}_{-i} \setminus \{\emptyset\}, SB_i(E) = \bigcap_{[h] \cap E \neq \emptyset} B_{i,h}(E).$$

In words,  $SB_i(E)$  is the event “player  $i$  would believe  $E$  conditional on every history that does not contradict  $E$ ”;<sup>37</sup> *e.g.*,  $SB_i([s_j])$  is the event “player  $i$  would

<sup>33</sup>The higher  $\beta$ , the more Bob believes that Ann’s choice to trust him is self-interested.

<sup>34</sup>This can happen in standard games too, but for different reasons and with more complex extensive forms (see, *e.g.*, Reny, 1992).

<sup>35</sup>See, *e.g.*, Battigalli & Bonanno (1999), Asheim (2005), and references therein.

<sup>36</sup>Here is an exact definition: for every event  $E = \bigcap_{i \in N} E_i$ ,  $E_i \in \mathcal{E}_i$ , let  $B(E) = \bigcap_{i \in N} B_{i,h^0} \left( \bigcap_{j \neq i} E_j \right)$ . Require that  $\omega \in R \cap \left( \bigcap_{k \geq 1} B^k(E) \right)$ , where  $B^k(E) = B^{k-1}(E)$ .

<sup>37</sup> $SB_i(\cdot)$  is not a monotone operator, and satisfies only a weak form of conjunctiveness  $[SB_i(E) \cap$

believe player  $j$  plays  $s_j$  at each history  $h$  allowed by  $s_j$ ".

We are interested in events of the form  $\text{SB}_i(R_{-i} \cap E)$ , where  $R_{-i} = \bigcap_{j \neq i} R_j$  is the event that  $i$ 's opponents are rational and  $E$  is either  $\Omega$  or some event concerning beliefs, and we consider assumptions like "everybody strongly believes that the opponents are rational." To write this concisely, we define a *mutual strong belief* operator. Let  $\mathcal{E}$  denote the collection of events of the form  $E = \bigcap_{i \in N} E_i$  ( $E_i \in \mathcal{E}_i$ ).

For example,  $R = \bigcap_{i \in N} R_i \in \mathcal{E}$ . For each  $E = \bigcap_{i \in N} E_i \in \mathcal{E}$ , the event "mutual strong

belief in  $E$ " is  $\text{SB}(E) = \bigcap_{i \in N} \text{SB}_i \left( \bigcap_{j \neq i} E_j \right)$ . Note that  $\text{SB}(E) \in \mathcal{E}$ .

We explore the consequences of the following assumptions:

- (0) each player is rational [=R],
- (1) mutual strong belief in (0) [=SB(R)],
- (2) mutual strong belief in (0) & (1) [=SB(R  $\cap$  SB(R))],
- (3) mutual strong belief in (0), (1) & (2) [=SB(R  $\cap$  SB(R  $\cap$  SB(R)))],

and so on.... Such assumptions are more easily expressed with formulas if we introduce an auxiliary 'correct strong belief' operator:

$$\forall E \in \mathcal{E}, \text{CSB}(E) = E \cap \text{SB}(E)$$

The conjunction of assumptions (0)-(k) corresponds to the event  $\text{CSB}^k(R)$ , where for any  $E \in \mathcal{E}$ ,  $\text{CSB}^0(E) = E$  and  $\text{CSB}^k(E) = \text{CSB}(\text{CSB}^{k-1}(E))$ .<sup>38</sup> Rationalizability is defined by considering the limit as  $k \rightarrow \infty$ :

**Definition 12** A state of the world  $\omega$  is rationalizable if  $\omega \in \bigcap_{k \geq 0} \text{CSB}^k(R)$ .

Battigalli & Siniscalchi (2002) show that the strategies consistent with event  $\text{CSB}^k(R)$  in standard games are those surviving the first  $k + 1$  steps of Pearce's (1984) extensive-form rationalizability procedure. This explains the terminology of Definition 12. To illustrate the concept, we note that it captures the forward induction solution of the Trust Game with guilt aversion (either  $\Gamma_2$  or  $\Gamma_3$ ). However, that conclusion requires only two layers of mutual correct strong belief: the solution obtains at all states  $\omega \in \text{CSB}^2(R)$ .

To illustrate the full power of Definition 12, we therefore analyze a *Generalized Trust Game with guilt aversion*, reminiscent of Ben-Porath & Dekel's (1992)

---

$\text{SB}_i(F) \subseteq \text{SB}_i(E \cap F)$ . For more on this, see Battigalli & Siniscalchi (2002).

<sup>38</sup>For example, (0) & (1) & (2) is  $R \cap \text{SB}(R) \cap \text{SB}(R \cap \text{SB}(R)) = \text{CSB}^2(R)$ .

money-burning game: Ann can either (evenly) distribute the total surplus of \$2 (action  $D$ ), or reinvest it in one out of  $L$  projects. Project  $\ell = 1, \dots, L$  yields  $\$2 \left(1 + \frac{\ell}{L}\right)$ , but Bob controls the distribution of the surplus and can either *Grab* or (evenly) *Share*. We let  $Trust_\ell$  denote the action of investing in project  $\ell$ , and  $\mathbf{Share}_\ell$  denote the conditional choice of sharing if Ann invests in project  $\ell$ . Let  $\alpha_\ell(\boldsymbol{\mu}_1) = \mu_1^1(\mathbf{Share}_\ell|h^0)$  and  $\beta_\ell(\boldsymbol{\mu}_2) = \int \alpha_\ell(\mu_1^1)\mu_2^2(d\alpha_\ell(\mu_1^1)|Trust_\ell)$ . As before we assume that Ann's utility is her material payoff, whereas Bob is averse to guilt. Applying the guilt formula of subsection 3.3, the players' utilities are given by

$$\begin{aligned} u_i(D) &= 1, \quad i = 1, 2, \\ u_i(Trust_\ell, Share) &= \left(1 + \frac{\ell}{L}\right), \quad i = 1, 2, \\ u_1(Trust_\ell, Grab) &= 0, \\ u_2(Trust_\ell, Grab) &= 2 \left(1 + \frac{\ell}{L}\right) - \theta_2 \alpha_\ell \left(1 + \frac{\ell}{L}\right), \end{aligned}$$

where  $\theta_2$  is Bob's sensitivity to guilt. Bob (strictly) prefers to share the yield of project  $\ell$  if and only if  $\theta_2 \beta_\ell > 1$ .

For  $L = 1$  and  $\theta_2 = \frac{5}{2}$  we obtain  $\Gamma_3$ , and the forward induction argument used to solve  $\Gamma_3$  (captured by 2 iterations of the CSB operator) works if and only if  $\theta_2 > 2$ . By contrast, when  $L > 1$  rationalizability yields the efficient sharing outcome also for much lower values of  $\theta_2$ :

**Proposition 13** *In the Generalized Trust Game with guilt aversion, if  $\theta_2 > 1 + \frac{1}{L}$  then, for every rationalizable state  $(s_1, \boldsymbol{\mu}_1, s_2, \boldsymbol{\mu}_2)$ ,  $s_1 = Trust_L$ ,  $s_2 = (\mathbf{Share}_\ell)_{\ell=1}^L$ ,  $\alpha_\ell(\boldsymbol{\mu}_1) = \beta_\ell(\boldsymbol{\mu}_2) = 1$  ( $\ell = 1, \dots, L$ ).*

**Proof.** Available on request. ■

The following theorem shows that our extension of Pearce's solution concept to psychological games is well behaved.

**Theorem 14** *If psychological utilities are continuous the set  $\bigcap_{k \geq 0} CSB^k(R)$  of rationalizable states is nonempty and compact.*

**Proof.** By definition

$$CSB^{k+1}(R) = CSB(CSB^k(R)) = CSB^k(R) \cap SB(CSB^k(R)) \subseteq CSB^k(R).$$

We prove by induction that each element  $\text{CSB}^k(R) = \bigcap_{\ell=0}^k \text{CSB}^\ell(R)$  of the nested sequence  $\{\text{CSB}^k(R)\}_{k \geq 0}$  is closed and nonempty. Lemma 3 implies  $\Omega$  is compact; thus, the closed subset  $\bigcap_{k \geq 0} \text{CSB}^k(R)$  is compact. Furthermore, the finite intersection property of compact spaces implies  $\bigcap_{k \geq 0} \text{CSB}^k(R) \neq \emptyset$ .

The inductive argument relies on the following three preliminary results, which are proved in the appendix.

**Lemma 15** *Correspondence  $r_i : \mathbf{M}_i \rightarrow S_i$  is nonempty valued. If  $u_i$  is also continuous,  $r_i$  has a closed graph and  $R_i$  is a nonempty closed set.*

**Lemma 16** *For every closed event  $E \in \mathcal{E}$ ,  $\text{SB}(E)$  is closed.*

**Lemma 17** *Let  $\{E^\ell\}_{\ell=0}^{\ell=k}$  be a decreasing sequence of nonempty events in  $\mathcal{E}$  ( $\emptyset \neq E^k \subseteq E^{k-1} \subseteq \dots \subseteq E^0$ ), then  $\bigcap_{\ell=0}^{\ell=k} \text{SB}(E^\ell)$  is also nonempty.*

For notational convenience let  $\text{CSB}^{-1}(E) = \Omega$ . We prove by induction that, for each  $k \geq 0$ ,  $\text{CSB}^k(R)$  is nonempty closed and can be expressed as

$$\text{CSB}^k(R) = R \cap \left( \bigcap_{\ell=-1}^{k-1} \text{SB}(\text{CSB}^\ell(R)) \right).$$

*Basis step.* The statement is true for  $k = 0$  because by Lemma 15  $\text{CSB}^0(R) = R$  is nonempty closed, and  $R$  can be expressed as

$$R = R \cap \Omega = R \cap \text{CSB}^{-1}(R)$$

*Inductive step.* Suppose the statement is true for each  $\ell = 0, \dots, k$ , then

$$\begin{aligned} \text{CSB}^{k+1}(R) &= \text{CSB}(\text{CSB}^k(R)) = \text{CSB}^k(R) \cap \text{SB}(\text{CSB}^k(R)) \\ &= R \cap \left( \bigcap_{\ell=-1}^{k-1} \text{SB}(\text{CSB}^\ell(R)) \right) \cap \text{SB}(\text{CSB}^k(R)) \\ &= R \cap \left( \bigcap_{\ell=-1}^k \text{SB}(\text{CSB}^\ell(R)) \right). \end{aligned}$$

By the inductive hypothesis each  $\text{CSB}^\ell(R)$  is nonempty and closed ( $\ell = 0, \dots, k$ ). By Lemma 16 also  $\text{SB}(\text{CSB}^\ell(R))$  is closed ( $\ell = 0, \dots, k$ ).  $R$  is also closed (Lemma 15). Hence  $\text{CSB}^{k+1}(R)$  is closed.  $\{\text{CSB}^\ell(R)\}_{\ell=0}^{\ell=k}$  is a decreasing sequence of nonempty events in  $\mathcal{E}$ . Therefore Lemma 17 implies that  $\bigcap_{\ell=-1}^k \text{SB}(\text{CSB}^\ell(R)) \neq \emptyset$ .

Pick any  $\omega = (s_i, \mu_i)_{i \in N} \in \bigcap_{\ell=-1}^k \text{SB}(\text{CSB}^\ell(R))$ . Since the latter is just an event about beliefs, modifying the strategies in  $\omega$  we obtain another state in the same event. By definition of  $R$ ,  $\prod_{i \in N} r_i(\mu_i) \times \{\mu_i\} \subseteq R$ . By Lemma 15,  $r_i(\mu_i) \neq \emptyset$ . We get

$$\emptyset \neq \prod_{i \in N} r_i(\mu_i) \times \{\mu_i\} \subseteq R \cap \left( \bigcap_{\ell=-1}^k \text{SB}(\text{CSB}^\ell(R)) \right).$$

Hence  $\text{CSB}^{k+1}(R) \neq \emptyset$ . This proves the inductive step, and the theorem. ■

## 6 Discussion and Extensions

In this section we compare our framework with GPS (6.1) and provide extensions concerning incomplete information (6.2), imperfect observability of past actions (6.3), dependence of utility on own strategy and dynamic (in)consistency (6.4).

### 6.1 Comparison with GPS

In section 2 we presented our framework as a generalization of GPS. This is not literally true. The reason is twofold. First, GPS allow for imperfect information and chance moves. As we show below, these complications can be included in our framework. Second, GPS allow for explicit randomization whereas we exclude it. *Prima facie*, this difference may seem immaterial. GPS assume players maximize expected (psychological) utility given beliefs, and in their framework there is no incentive to randomize. It might seem that the only role played by randomization is to guarantee the existence of equilibrium, a result we obtain by looking at equilibrium in beliefs. However, unlike standard games, in psychological games there may be a difference (to a player's utility) between a belief assigning probability one to a randomized choice that, say, picks  $a$  or  $b$  with probability  $\frac{1}{2}$ , and the belief that assigns probability  $\frac{1}{2}$  to each of  $a$  and  $b$ . These beliefs are equivalent if psychological utility functions satisfy a linearity property. In most examples/applications of psychological games we are aware of this holds.

Now look at the version of GPS that *is* a special case of our framework: psychological games with utilities of the form  $u_i : Z \times \overline{\mathbf{M}}_i \rightarrow \mathbb{R}$ , where  $\overline{\mathbf{M}}_i$  is the space of infinite hierarchies of *initial* beliefs of  $i$ , and first-order beliefs are probability measures over pure strategies of  $i$ 's opponents. How much is lost by restricting the analysis to such games? We have argued that many interesting phenomena such as sequential reciprocity, psychological forward induction, and regret cannot be analyzed. However, we can prove a partial equivalence result. Suppose the initial

beliefs of others enter the utility  $u_i : Z \times \prod_{j \in N} \overline{\mathbf{M}}_j \rightarrow \mathbb{R}$ . Then there is a psychological game with utilities  $\overline{u}_i : Z \times \overline{\mathbf{M}}_i \rightarrow \mathbb{R}$  that has the same sequential equilibrium assessments as the former game.<sup>39</sup> This does not mean that in this class of games conditional higher-order beliefs are immaterial. First, the equivalence result only concerns sequential equilibria, and we argued that the non-equilibrium analysis of psychological games is important. Second, our very definition of sequential equilibrium makes essential use of conditional beliefs.<sup>40</sup>

## 6.2 Incomplete information

Unless one models interaction within a family or amongst friends, it is probably not realistic to assume that players know one another's psychological propensities. Many of our examples can be criticized on that ground. For example, in the analysis of  $\Gamma_2$  (or  $\Gamma_3$ ) we assumed that Ann knows that Bob's sensitivity to guilt is  $\theta_2 = \frac{5}{2}$ , which may be a stretch.<sup>41</sup> Another reason to allow for incomplete information is that a player may care about the beliefs of others about some of his characteristics, which are not common knowledge, as in the models of Bernheim and Dufwenberg & Lundholm.

In order to extend the analysis of psychological games to include incomplete information, let  $\theta = (\theta_0, \theta_1, \dots, \theta_n)$  denote a vector of parameters that summarize all the payoff-relevant aspects of the game that are not common knowledge;  $\theta_i$  is a component known to player  $i$  only (such as his ability, or his sensitivity to certain psychological motivations), nobody knows  $\theta_0$ . It is common knowledge that  $\theta$  belongs to a parameter space  $\Theta = \Theta_0 \times \Theta_1 \times \dots \times \Theta_n$ . Elements of  $\Theta$  are called *states of Nature*. Assume  $\Theta$  is a compact Polish space. Also assume for simplicity that players do not get more refined information about the state of Nature as play unfolds, they only observe the actions chosen in previous stages.

<sup>39</sup>The intuition is relatively simple: each initial belief hierarchy  $\overline{\boldsymbol{\mu}}_i$  induces a probability measure  $\overline{f}_i(\overline{\boldsymbol{\mu}}_i) \in \Delta(S_{-i} \times \overline{\mathbf{M}}_{-i})$  which can be used to compute an expectation  $\overline{u}_i(z, \overline{\boldsymbol{\mu}}_i)$  of  $u_i(z, \overline{\boldsymbol{\mu}}_i, \cdot)$ . Since in a consistent assessment there is 'common knowledge' of the hierarchical beliefs, no observation will make the players change their mind about the initial beliefs of the opponents, hence for any consistent assessment  $u_i$  and  $\overline{u}_i$  have the same set of maximizing actions at each history. (If the game has only one stage  $u_i$  and  $\overline{u}_i$  are fully equivalent, *i.e.*, they have the same best response correspondences.)

<sup>40</sup>If conditional beliefs were not in the language we would have to use an indirect approach similar to the one adopted by GPS: First define what a psychological Nash equilibrium is using the *ex ante* versions of Definitions 6 and 8. Then stipulate that a Nash equilibrium profile  $\overline{\boldsymbol{\mu}}$  is a sequential equilibrium if there is a behavioral strategy profile  $\sigma$  that is a sequential equilibrium of the standard game with payoff functions  $u_i(\cdot, \overline{\boldsymbol{\mu}})$ , and is such that each  $\text{marg}_{S_i} \overline{\boldsymbol{\mu}}_j^1$  ( $j \neq i$ ) is derived from  $\sigma_i$  *via* Kuhn's transformation.

<sup>41</sup>Ample evidence in psychology suggests emotional sensitivities differ among people. See Krohne (2003) for a general discussion, and Tangney (1995) on guilt specifically.

It is relatively easy to generalize our construction of the belief space in order to include beliefs about the state of Nature: replace  $X_{-i}^0 = S_{-i}$  with  $X_{-i}^0 = S_{-i} \times \Theta_{-i}$  in the construction of subsection 3.2, where  $\Theta_{-i} = \Theta_0 \times \dots \times \Theta_{i-1} \times \Theta_{i+1} \times \dots \times \Theta_n$ . Conditioning events for first-order beliefs now have the form  $F = S_{-i}(h) \times \Theta_{-i}$  ( $h \in H$ ). Let  $X_{-i}^{k-1}$  be the space of  $(k-1)$ -order uncertainty for player  $i$ ; then we obtain the set of  $k$ -order cps'  $\Delta^H(X_{-i}^{k-1})$ , and the  $k$ -order uncertainty space  $X_{-i}^k = X_{-i}^{k-1} \times \prod_{j \neq i} \Delta^H(X_{-j}^{k-1})$ . Lemmata 2 and 3 are easily extended to this case. Therefore we obtain, for each  $i \in N$ , the space  $\mathbf{M}_i$  of infinite hierarchies of cps' consistent with collective coherency, which is a compact Polish space homeomorphic to  $\Delta^H(S_{-i} \times \Theta_{-i} \times \mathbf{M}_{-i})$ .<sup>42</sup>

This is all we need to define the domain of psychological payoff functions, but it need not exhaust the description of the psychological game. Since states of Nature are exogenous, also players' hierarchies of *initial* beliefs about the state of Nature are exogenous,<sup>43</sup> and the model may specify assumptions about such exogenous beliefs. For example, one may assume that beliefs about  $\theta$  are derived from a common prior  $\rho \in \Delta(\Theta)$  and that this is common knowledge.<sup>44</sup> More sophisticated assumptions are allowed by Harsanyi's implicit representation of belief hierarchies by means of a  $\Theta$ -based type space (cf. Harsanyi, 1967-68; Mertens & Zamir, 1985; Brandenburger & Dekel, 1993). Alternatively, assumptions about exogenous beliefs may be stated explicitly. Whatever these assumptions may be, they identify subspaces of hierarchies of cps'  $\hat{\mathbf{M}}_i \subseteq \mathbf{M}_i$ ,  $i = N$ , which form the basis for the strategic analysis of the game. The analysis of rationalizability can be quite easily extended to this more general framework.<sup>45</sup> Sequential equilibrium requires more care because the extension of the definition of consistency to general games of incomplete information is not obvious.

With this extended framework, we can regard (appropriately discretized versions of) the models of conformity (Bernheim) and social respect (Dufwenberg & Lundholm) as psychological games with incomplete information. These models have a non-standard signaling game structure. There is only one active player, player 1, who has private information  $\theta$  and chooses action  $a$ . Player 2 ('society')

---

<sup>42</sup>See Battigalli & Siniscalchi (1999).

<sup>43</sup>Posterior beliefs about the state of Nature are endogenous, because they are derived (by conditioning on histories) from joint beliefs about strategies (and beliefs of others) and about the state of Nature.

<sup>44</sup>Even in this simple case, we should distinguish the incomplete information situation from one where there is asymmetric information about chance moves, especially in the rationalizability analysis. On chance moves see the next subsection.

<sup>45</sup>A state of player  $i$  is a triple  $(s_i, \theta_i, \mu_i)$ ; the definition of rationality is almost the same as in section 5, except that player  $i$  takes into account her knowledge  $\theta_i$  of the state of Nature. See Battigalli & Siniscalchi (2002).

makes no choice, but observes  $a$  and makes inferences about  $\theta$ . The payoff function of player 1 has the form  $u_1 : A \times \Theta \times \Delta^H(A \times \Theta) \rightarrow \mathbb{R}$  (where  $H = \{h^0\} \cup A$ ). More specifically, there is a valuation function  $v_1 : A \times \Theta \times \Delta(\Theta) \rightarrow \mathbb{R}$  such that

$$u_1(a, \theta, \mu_2) = v_1(a, \theta, \mu_2(\cdot|a)).$$

This means that player 1 cares about the beliefs player 2 will hold about private information  $\theta$  conditional on his action  $a$ . (Note that this is an instance where *terminal* beliefs affect payoffs.)

In Caplin & Leahy's (2004) model an information providing doctor is concerned about the anxiety of the patient, which in turn depends on her posterior beliefs about her health; this may induce the doctor not to reveal information. Caplin & Leahy note that their game does not fit GPS' framework and take a detour to get around this. But it can be shown that their game fits our framework.

### 6.3 Imperfectly observable actions and chance moves

We chose to focus on games with observable actions and no chance moves for the sake of simplicity. But our concepts and results carry over to the more general case of games where past actions need not be perfectly observed and chance may play a role (as in GPS). Let  $N = \{0, 1, \dots, n\}$  where index 0 denotes the chance player, and let  $\mathbf{H}_i$  be the partition of the set of histories  $H$  into information sets of player  $i$  ( $i \neq 0$ ).<sup>46</sup> Assume perfect recall holds. Then the set of strategy profiles consistent with any information set  $\mathbf{h}_i \in \mathbf{H}_i$  must have the form  $S(\mathbf{h}_i) = S_i(\mathbf{h}_i) \times S_{-i}(\mathbf{h}_i)$ . We have to consider, for the first-order beliefs of player  $i$ , the collection of conditioning events  $\{F_i : F_i = S_{-i}(\mathbf{h}_i), \mathbf{h}_i \in \mathbf{H}_i\}$ . Let  $X_{-i}^{k-1}$  be the space of  $(k-1)$ -order uncertainty for player  $i$ ; then we obtain the set of  $k$ -order cps'  $\Delta^{\mathbf{H}_i}(X_{-i}^{k-1})$ , and the  $k$ -order uncertainty space  $X_{-i}^k = X_{-i}^{k-1} \times \prod_{j \neq i, 0} \Delta^{\mathbf{H}_j}(X_{-j}^{k-1})$ . The resulting set of infinite hierarchies of cps'  $\mathbf{M}_i$  is homeomorphic to  $\Delta^{\mathbf{H}_i}(S_{-i} \times \mathbf{M}_{-i})$ . (Restricted sets of hierarchies  $\hat{\mathbf{M}}_i$  reflect 'common knowledge' of the probabilities of chance moves.) As in the case of incomplete information, the analysis of rationalizability is easily extended, while the definition of consistency in the sequential equilibrium analysis requires more care.<sup>47</sup>

<sup>46</sup>Note that also terminal histories are partitioned into information sets because terminal beliefs are allowed to play a role.

<sup>47</sup>The easiest way to define consistency (although not the most transparent) is to replace (a) and (b) of Definition 6 with a topological condition similar to the one originally used by Kreps & Wilson. Similar considerations apply to games with incomplete information.

## 6.4 Own-strategy dependence and dynamic (in)consistency

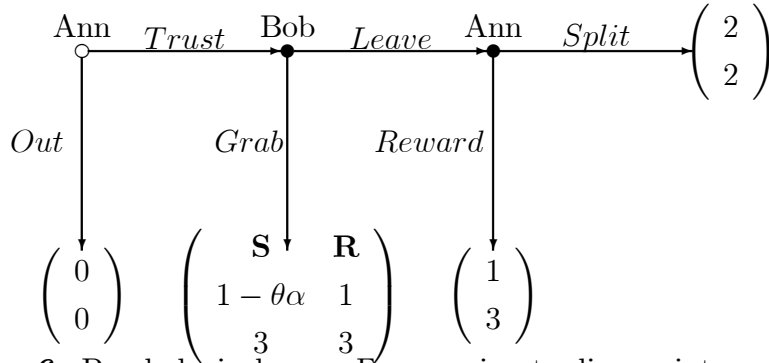
We have argued that it is natural to let a player's utility depend directly on the other players' strategies, but we have so far assumed that it does not depend directly on his own strategy. This assumption allowed us to apply standard dynamic programming techniques and prove Theorem 14. In this subsection we show that allowing for own-strategy *dependence* is natural and gives rise to an interesting form of dynamic inconsistency.

Consider the following version of the Trust Game: Ann can either trust Bob or opt out. If she opts out no surplus is created; if she trusts Bob the total surplus is \$4 and Bob can either grab \$3 or let Ann allocate the surplus (action *Leave*). If Bob leaves the allocation to Ann she can either split the surplus or reward Bob, giving him the \$3 he could have grabbed.

Now assume that if Ann gets less money than she expected she feels disappointed and that the anticipation of this negative feeling affects her decisions. This is captured by the following utility function:

$$u_1(z, \boldsymbol{\mu}, s) = \pi_1(z) - \theta D_1(z, \mu_1^1(\cdot|h^0), s_1)$$

where  $\theta$  is a psychological sensitivity parameter and  $D_j(z, \mu_j^1(\cdot|h^0), s_j)$  is the disappointment function defined in subsection 3.3.  $\Gamma_6$  builds on this function, and thereby turns out to exhibit own-strategy dependent utility for Ann. (For simplicity, we let  $u_2(z, \boldsymbol{\mu}, s) = \pi_2(z)$ ).



**Figure 6.** Psychological game  $\Gamma_6$ : aversion to disappointment.

The utility assigned by Ann to terminal history  $(Trust, Grab)$  depends on her initial belief  $\alpha = \mu_1^1(\mathbf{Leave}|h^0)$ , and on how she plans to behave if Bob leaves for her to allocate the surplus. The own-strategy dependence arises because Ann cannot feel disappointed if she plans to reward Bob (column **R**), since in this case the resulting material payoff is 1 regardless of what Bob chooses. For Ann to be dissatisfied at terminal history  $(Trust, Grab)$  requires that she plans to split

and that  $\alpha > 0$ . In this case she (initially) expects a material payoff which is larger than her material payoff at history  $(Trust, Grab)$   $[(1 - \alpha) \times 1 + \alpha \times 2 > 1]$ . The disappointment yields utility  $(1 - \theta\alpha)$  (column **S**). The expected utility of plan  $(Trust, \mathbf{Split})$  is thus  $(1 - \alpha) \times (1 - \theta\alpha) + \alpha \times 2$ , which could be lower than 1 (in fact, even lower than 0) if  $\theta$  is large enough. In this case, the *ex ante* expected utility maximizing plan is  $(Trust, \mathbf{Reward})$  (which prevents disappointment and yields 1). However,  $(Trust, \mathbf{Reward})$  is not dynamically consistent because the best choice after history  $(Trust, Leave)$  is to split. As a result there is no strategy that maximizes Ann's expected utility at the beginning of the game and also at history  $(Trust, Leave)$ . (A similar kind of dynamic inconsistency arises in Caplin & Leahy's (2001) theory of psychological expected utility, which can be shown to be consistent with our extended framework; see the appendix.)

Since Ann cannot commit to reward Bob, she should initially maximize under the constraint that she would split in the endgame; the resulting plan is *Out*. This kind of 'consistent planning' (Strotz, 1956) can be formally represented within a multi-self approach by requiring that Ann's strategy be immune to one-shot deviations.<sup>48</sup> However, it can be shown that even this weaker rationality condition may be impossible to satisfy, unless we allow for uncertainty about one's own strategy.<sup>49</sup>

Besides own-strategy dependence, a more direct way to allow for dynamic inconsistencies is to adopt a multi-self approach and model a player's preferences with an array of 'local' utility functions  $(u_{i,h} : Z \times \mathbf{M} \times S \rightarrow \mathbb{R})_{h \in H \setminus Z}$ . The sequential equilibrium analysis of section 4 applies to this extended framework almost *verbatim*.

This formulation is relevant to reciprocity theory. We have already seen how our basic framework could reproduce Dufwenberg & Kirchsteiger's theory in an example ( $\Gamma_6$ ). However, to handle general games one needs a multi-selves approach, and it is then possible to (essentially) reformulate Dufwenberg & Kirchsteiger's model (details are available on request).

---

<sup>48</sup>This is true in the example, not in general. Since 'later selves' may be indifferent, immunity to one-shot deviations is necessary but not sufficient for consistent planning.

<sup>49</sup>This can be done, at the cost of additional complexity, within a richer framework where player  $i$ 's first-order beliefs are defined over  $S$  rather than  $S_{-i}$  (cf. Battigalli & Siniscalchi, 1999).

## 7 Concluding remarks

Psychologists are more likely than economists to discuss emotions and social rewards, while economists are more likely to use mathematical models of incentives in interactive situations. We offer a synthesis, a mathematical framework that accommodates belief-dependent motivation in game theoretic contexts and captures how certain emotions and social values affect decision making and strategic reasoning. Camerer, Loewenstein & Prelec (2005; pp. 10, 55), in a discussion of how neuroscience can inform economics, talk about “incremental” research, in which “psychological evidence suggests functional forms” which “enhance the realism of existing [economic] models”. Our approach may be seen as providing tools that can facilitate such an approach.

We propose that there are a variety of interesting psychological phenomena waiting to be analytically explored. In his survey paper on “Emotions and Economic Theory”, Elster (1998) argues that a key characteristic of emotions is that “they are triggered by beliefs” (p. 49). He discusses, *i.a.*, anger, hatred, guilt, shame, pride, admiration, regret, rejoicing, disappointment, elation, fear, hope, joy, grief, envy, malice, indignation, jealousy, surprise, boredom, sexual desire, enjoyment, worry, and frustration. Some of his examples involve higher-order beliefs. He asks (p. 48): “[H]ow can emotions help us explain behavior for which good explanations seem to be lacking?” The framework we develop in this paper may be useful for providing answers.

Loewenstein, Weber, Hsee & Welch (2001) distinguish between “anticipated” emotions which “are expected to occur when outcomes are experienced” (p. 268), and “anticipatory” emotions which “are immediate visceral reactions” that “often drive behavior” (p. 267). Our framework can address both kinds. For example, our analysis of guilt (*e.g.* in  $\Gamma_2$  and  $\Gamma_3$ ) concerns an anticipated emotion, while Caplin & Leahy’s (2001, 2004) analysis of anxiety (which one can show fits our framework; cf. 6.2, 6.4 and the appendix) concerns anticipatory emotions. Our framework can also accommodate concern for the feelings of others, like in Caplin & Leahy (2004) where a doctor cares about the anxiety level of his patient.

In closing, we note a limitation which is buried in the solution concepts we develop. They presume (one way or another, and if taken at face value) that players are rational. Of course, lots of psychological research indicates how actual people suffer from cognitive bounds and illusions. While this should be noted, we do not believe the matter is worth any lost sleep. The idea of belief-dependent motivation is not necessarily tied to unbounded rationality. Modeling bounded

rationality is beyond the scope of our paper, but if that task were tackled the definition of a psychological game proposed in this paper would seem relevant even if our solution concepts are not.

## 8 Appendix

We start with some preliminaries about rationality and backward induction on belief-induced decision trees, and then prove Lemmata 15, 16 and 17. For any fixed hierarchy of cps'  $\mu_i$ , we obtain a well defined decision tree that can be solved by backward induction: define value functions  $V_{\mu_i} : H \rightarrow \mathbb{R}$  and  $\bar{V}_{\mu_i} : (H \setminus Z) \times A_i \rightarrow \mathbb{R}$  as follows

- For terminal histories  $z \in Z$ , let

$$V_{\mu_i}(z) = \int_{S_{-i} \times \mathbf{M}_{-i}} u_i(z, \boldsymbol{\mu}_i, \boldsymbol{\mu}_{-i}, s_{-i}) f_{i,z}(\boldsymbol{\mu}_i) (ds_{-i}, d\boldsymbol{\mu}_{-i}).$$

- Assuming that  $V_{\mu_i}(h, a)$  has been defined for the immediate successors  $(h, a)$  of history  $h$ , let

$$\bar{V}_{\mu_i}(h, a_i) = \sum_{a_{-i} \in A_{-i}(h)} \mu_i^1(S_{-i}(h, a_{-i}) | h) V_{\mu_i}(h, (a_i, a_{-i}));$$

for each  $a_i \in A_i(h)$ ; then  $V_{\mu_i}(h)$  is defined as

$$V_{\mu_i}(h) = \max_{a_i \in A_i(h)} \bar{V}_{\mu_i}(h, a_i).$$

Recall that, for any strategy  $s_i \in S_i$ ,  $H_i(s_i) = \{h \in H \setminus Z : s_i \in S_i(h)\}$  denotes the set of histories allowed by  $s_i$ . The proof of the following result is available by request:

**Lemma 18** *The sequential best reply correspondence  $r_i : M_i \rightarrow S_i$  can be characterized as follows*

$$r_i(\boldsymbol{\mu}_i) = \left\{ s_i : \forall h \in H(s_i), s_{i,h} \in \arg \max_{a_i \in A_i(h)} \bar{V}_{\mu_i}(h, a_i) \right\}.$$

### Proof of Lemma 15

By Lemma 18  $r_i(\mu_i) = \{s_i : \forall h \in H(s_i), s_{i,h} \in \arg \max_{a_i \in A_i(h)} \bar{V}_{\mu_i}(h, a_i)\}$ . Clearly, the RHS is nonempty. Therefore  $r_i(\cdot)$  is nonempty-valued and  $R_i$  is nonempty.

The belief function  $f_i$  is continuous (Lemma 3). If  $u_i$  is also continuous, then  $E_{s_i, \mu_i}[u_i|h]$  is continuous (in  $\mu_i$ ), which implies that  $R_i$  is closed. ■

### Proof of Lemma 16

We must show that for every closed event  $E \in E_{-i}$ ,  $\text{SB}_i(E)$  is closed.  $\text{SB}_i(\emptyset) = \emptyset$ , a closed set, by definition. Suppose that  $E = \Omega_i \times E_{-i}$  where  $E_{-i}$  is nonempty and closed. Recall that  $\text{SB}_i(E) = \bigcap_{h:[h] \cap E \neq \emptyset} \text{B}_{i,h}(E)$ . For each  $h$ ,

$$\text{B}_{i,h}(E) = S_i \times f_{i,h}^{-1}(\Delta(E_{-i} \cap (S_{-i}(h) \times M_{-i}))) \times \Omega_{-i},$$

where for any measurable space  $X$  and any  $F \subseteq X$  we let  $\Delta(F)$  denote the set of probability measures on  $X$  that assign probability one to  $F$ . Note that if  $F$  is closed,  $\Delta(F)$  is also closed. The coordinate function  $f_{i,h} : M_i \rightarrow \Delta(\Omega_{-i})$  is continuous and  $M_{-i}$  is closed (Lemma 3); hence  $E_{-i} \cap (S_{-i}(h) \times M_{-i})$ ,  $\Delta(E_{-i} \cap (S_{-i}(h) \times M_{-i}))$  and  $f_{i,h}^{-1}(\Delta(E_{-i} \cap (S_{-i}(h) \times M_{-i})))$  are closed. It follows that  $\text{B}_{i,h}(E)$  ( $h \in H$ ) and  $\text{SB}_i(E)$  are closed. ■

### Proof of Lemma 17

Let  $\{E^\ell\}_{\ell=0}^{\ell=k}$  be a decreasing sequence of nonempty events in  $E$  ( $\emptyset \neq E^k \subseteq E^{k-1} \subseteq \dots \subseteq E^0$ ); we show that  $\bigcap_{\ell=0}^{\ell=k} \text{SB}(E^\ell)$  is also nonempty. For each  $\ell$  and  $i$ ,  $E^\ell \in E$  can be written  $E^\ell = E_i^\ell \times E_{-i}^\ell$ , where  $E_{-i}^\ell \subseteq \Omega_{-i}$ , and by definition of  $\text{SB}(\cdot)$

$$\bigcap_{\ell=0}^{\ell=k} \text{SB}(E^\ell) = \bigcap_{i \in N} \bigcap_{\ell=0}^{\ell=k} \text{SB}_i(\Omega_i \times E_{-i}^\ell).$$

Therefore we must show that  $\bigcap_{\ell=0}^{\ell=k} \text{SB}_i(\Omega_i \times E_{-i}^\ell) \neq \emptyset$  ( $i \in N$ ). Let  $\Delta^H(\Omega_{-i}; E_{-i}^\ell)$  denote the set of cps'  $\mu \in \Delta^H(\Omega_{-i})$  such that  $\mu(E_{-i}^\ell|h) = 1$  for each  $h$  such that  $E_{-i}^\ell \cap (S_{-i}(h) \times M_{-i}) \neq \emptyset$ . Note that

$$\bigcap_{\ell=0}^{\ell=k} \text{SB}_i(\Omega_i \times E_{-i}^\ell) = S_i \times f_i^{-1} \left( \bigcap_{\ell=0}^{\ell=k} \Delta^H(\Omega_{-i}; E_{-i}^\ell) \right) \times \Omega_{-i}.$$

We show below that  $\bigcap_{\ell=0}^{\ell=k} \Delta^H(\Omega_{-i}; E_{-i}^\ell) \neq \emptyset$ . Since  $f_i$  is onto (Lemma 3), it follows that  $f_i^{-1} \left( \bigcap_{\ell=0}^{\ell=k} \Delta^H(\Omega_{-i}; E_{-i}^\ell) \right) \neq \emptyset$ . Hence  $\bigcap_{\ell=0}^{\ell=k} \text{SB}_i(\Omega_i \times E_{-i}^\ell) \neq \emptyset$ .

We show that  $\bigcap_{\ell=0}^{\ell=k} \Delta^H(\Omega_{-i}; E_{-i}^\ell) \neq \emptyset$  with a recursive construction. Say that  $h$  is ‘reached’ by probability measure  $\nu \in \Delta(\Omega_{-i})$  if  $\nu(S_{-i}(h) \times M_{-i}) > 0$ . Note that if  $h$  is reached by  $\nu$ , every predecessor of  $h$  is also reached by  $\nu$ . Say that

$\mu(\cdot|h)$  is ‘derived’ from  $\nu$ , where  $\nu$  reaches  $h$ , if for every Borel set  $F_{-i} \subseteq \Omega_{-i}$

$$\mu(F_{-i}|h) = \frac{\nu(F_{-i} \cap (S_{-i}(h) \times \mathbf{M}_{-i}))}{\nu(S_{-i}(h) \times \mathbf{M}_{-i})}.$$

Pick any probability measure  $\nu$  in the (nonempty) set  $\Delta(E_{-i}^k)$ . For each  $h$  reached by  $\nu$  let  $\mu(\cdot|h)$  be derived from  $\nu$ . Thus,  $\mu(\cdot|h)$  has been defined for a nonempty set of histories closed w.r.t. precedence (that is, if  $h$  is in the set every predecessor of  $h$  is in the set), the set is nonempty because it contains the initial history  $h^0$ . Now suppose that  $\mu(\cdot|h)$  has been defined for some set of histories  $\hat{H}$  closed w.r.t. precedence. If  $\hat{H} \neq H$ , for each  $h \in H \setminus \hat{H}$  such that the immediate predecessor of  $h$  belongs to  $\hat{H}$ , pick a probability measure  $\nu_h$  in the set  $\Delta(E_{-i}^{\ell(h)} \cap (S_{-i}(h) \times M_{-i}))$ , where  $\ell(h)$  is the highest index  $\ell \in \{-1, 0, \dots, k\}$  such that  $E_{-i}^{\ell} \cap (S_{-i}(h) \times M_{-i}) \neq \emptyset$ , and by convention we let  $E^{-1} = \Omega_{-i}$ . Let  $\mu(\cdot|h')$  be derived from  $\nu_h$  whenever  $h'$  weakly follows  $h$  and is reached by  $\nu_h$ . Now  $\mu(\cdot|h)$  is defined for a set of histories  $\hat{H}'$  closed under the precedence relation and strictly larger than  $\hat{H}$ . Proceed in this way until the whole  $H$  is covered. We claim that the resulting vector of probability measures  $(\mu(\cdot|h))_{h \in H}$  is a cps  $\mu \in \bigcap_{\ell=0}^{\ell=k} \Delta^H(\Omega_{-i}; E_{-i}^{\ell})$ .

To see that  $(\mu(\cdot|h))_{h \in H}$  is a cps we only have to check that the ‘chain rule’ (3) in Definition 1 holds. Suppose that  $h$  precedes  $h'$ . To write formulas more transparently, let  $C = S_{-i}(h) \times M_{-i}$ ,  $C'_{-i} = S_{-i}(h') \times M_{-i}$ ,  $\mu(\cdot|h) = \mu(\cdot|C_{-i})$ ,  $\mu(\cdot|h') = \mu(\cdot|C'_{-i})$ . Since  $h$  precedes  $h'$ ,  $S_{-i}(h') \subseteq S_{-i}(h)$ , hence  $C'_{-i} \subseteq C_{-i}$ . If  $h'$  is not reached by  $\mu(\cdot|C_{-i})$  then (3) holds trivially as  $0 = 0$ . If  $h'$  is reached by  $\mu(\cdot|C_{-i})$ , then  $\mu(\cdot|C_{-i})$  and  $\mu(\cdot|C'_{-i})$  are both derived from the same measure – say  $\nu \in \Delta(\Omega_{-i})$  – reaching both  $h$  and  $h'$ ; thus, for every Borel set  $F_{-i} \subseteq C'_{-i}$

$$\mu(F_{-i}|C_{-i}) = \frac{\nu(F_{-i})}{\nu(C_{-i})} = \frac{\nu(F_{-i})}{\nu(C'_{-i})} \frac{\nu(C'_{-i})}{\nu(C_{-i})} = \mu(F_{-i}|C'_{-i}) \mu(C'_{-i}|C_{-i}).$$

To see that  $\mu \in \bigcap_{\ell=0}^{\ell=k} \Delta^H(\Omega_{-i}; E_{-i}^{\ell})$ , note that by construction  $\mu(E^{\ell(h)}|h) = 1$  for all  $h \in H$ . Suppose that, for any index  $\ell \in \{0, \dots, k\}$  and any  $h \in H$ ,  $E_{-i}^{\ell} \cap (S_{-i}(h) \times M_{-i}) \neq \emptyset$ . Then  $\ell(h) \geq \ell$  and  $\mu(E^{\ell}|h) \geq \mu(E^{\ell(h)}|h) = 1$ ; hence  $\mu(E^{\ell}|h) = 1$  as desired. ■

### Anticipated and Anticipatory Feelings

We describe here a subclass of psychological games that highlight the effect of anticipated as well as anticipatory feelings (cf. Loewenstein et al 2001, and our discussion in section 7), including a concern for others who harbor such feelings. Consider a  $T$ -stage game with observable actions and player set  $\{0, 1, \dots, n\}$ ,

where 0 is Chance. At the end of each stage  $t$  an outcome  $o^t = O^t(h^t)$  realizes. Outcome  $o^t$  and  $i$ 's past and current beliefs give rise to a psychological state  $p_i^t = P_i^t(o^t, \rho_i^{0+}, \dots, \rho_i^{t+})$ , where  $\rho_i^{\tau+}$  is  $i$ 's probability measure over  $(o^{\tau+1}, \dots, o^T)$  computed at the end of stage  $\tau \leq t$ .<sup>50</sup> Assume that  $i$  maximizes the expectation of  $\sum_{t=1}^T v_i^t(p_1^t, \dots, p_n^t)$ , where  $v_i^t(p_1, \dots, p_n)$  is a one-stage utility function that captures the impact of stage- $t$  outcome and of anticipatory feelings (such as anxiety).

A pair  $(\mu_i^1, s_i) \in \Delta^H(S_{-i}) \times S_i$  induces an end-of-stage- $t$  conditional probability measure over future outcomes  $o^{t+1}, \dots, o^T$ , denoted by  $\rho_i^{t+}(\cdot | h^t; \mu_i^1, s_i)$ . Let  $h^t(z)$  denote the prefix of length  $t$  of terminal history  $z$ . The psychological game utility function of  $i = 1, \dots, n$  is

$$u_i(z, \boldsymbol{\mu}, s) = \sum_{t=1}^T v_i^t((P_j^t(O^t(h^t(z))), \rho_j^{0+}(\cdot | h^t(z); \mu_j^1, s_j), \dots, \rho_j^{t+}(\cdot | h^t(z); \mu_j^1, s_j))_{j=1}^n).$$

The expectation of  $u_i$  is computed by means of the second-order beliefs  $\mu_i^2$ .

For example, suppose  $v_i^t \equiv 0$  for each  $t < T$ . Then only anticipated feelings matter, as in the models of disappointment aversion and guilt aversion (in these models  $p_i^T$  depends only on  $\pi_i$  and  $\rho_i^{0+}$ ). On the other hand, in the psychological expected utility theory of Caplin & Leahy (2001) anticipatory feelings affect utility and behavior via  $v_i^t$  ( $t < T$ ) which depends only on  $p_i^t$ , which in turn depends only on  $o^t$  and  $\rho_i^{t+}$ . In Caplin & Leahy (2004)  $v_i^t$  depends also on  $p_j^t$ ,  $j \neq i$ .

## References

- [1] ASHEIM, G (2005): *The Consistent Preferences Approach to Deductive Reasoning in Games*. Springer (Due: September 2005)
- [2] AUMANN, R.J. AND A. BRANDENBURGER (1995): "Epistemic Conditions for Nash Equilibrium," *Econometrica*, 63, 1161-1180.
- [3] BACHARACH, M., G. GUERRA AND D.J. ZIZZO (2001): "Is Trust Self-Fulfilling? An Experimental Study," mimeo.
- [4] BATTIGALLI, P. (1996): "Strategic Independence and Perfect Bayesian Equilibria," *Journal of Economic Theory*, 70, 201-234.

---

<sup>50</sup>  $\rho_i^{0+}$  is  $i$ 's prior belief. We assume for notational convenience that  $\rho_i^{T+}$  assigns probability 1 to an exogenously given outcome  $\bar{o}^{T+1}$ .

- [5] BATTIGALLI, P. AND G. BONANNO (1999): “Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory,” *Research in Economics (Ricerche Economiche)*, 53, 149-226.
- [6] BATTIGALLI, P. AND M. SINISCALCHI (1999): “Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games,” *Journal of Economic Theory*, 88, 188-230.
- [7] — (2002): “Strong Belief and Forward Induction Reasoning,” *Journal of Economic Theory*, 106, 356-391.
- [8] BELL, D.E. (1982): “Regret in Decision Making under Uncertainty,” *Operations Research*, 30, 961-981.
- [9] BEN-PORATH, E. AND E. DEKEL (1992): “Signaling Future Actions and the Potential for Sacrifice,” *Journal of Economic Theory*, 57, 36-51.
- [10] BERNHEIM, D. (1984): “Rationalizable Strategic Behavior,” *Econometrica*, 52, 1007-1028.
- [11] — (1994): “A Theory of Conformity,” *Journal of Political Economy*, 102, 841-77.
- [12] BRANDENBURGER, A. AND E. DEKEL (1993): “Hierarchies of Beliefs and Common Knowledge,” *Journal of Economic Theory* 59, 189-198.
- [13] CAMERER, C., G. LOEWENSTEIN AND D. PRELEC (2005): “Neuroeconomics: How Neuroscience Can Inform Economics,” *Journal of Economic Literature*, 43, 9-64.
- [14] CAPLIN, A. (2003): “Fear as a Policy Instrument: Economic and Psychological Perspectives on Intertemporal Choice.” in Loewenstein, Read and Baumeister, eds, *Time and Decision*. New York: Russell Sage Foundation.
- [15] CAPLIN, A. AND K. ELIAZ (2003): “AIDS Policy and Psychology: a Mechanism design Approach,” *RAND Journal of Economics*, 34, 631-646.
- [16] CAPLIN, A. AND J. LEAHY (2001): “Psychological Expected Utility Theory and Anticipatory Feelings,” *Quarterly Journal of Economics*, 116, 55-79.
- [17] — (2004): “The Supply of Information by a Concerned Expert,” *Economic Journal*, 114, 487-505.

- [18] CHARNES, G. AND M. DUFWENBERG (2004): “Promises and Partnership”, mimeo, UCSB and University of Arizona.
- [19] DUFWENBERG, M. (1995): “Time-Consistent Wedlock with Endogenous Trust”, in Doctoral Dissertation, Uppsala University.
- [20] — (2002): “Marital Investment, Time Consistency and Emotions,” *Journal of Economic Behavior and Organization*, 48, 57-69.
- [21] DUFWENBERG, M. AND U. GNEEZY (2000): “Measuring beliefs in an experimental lost wallet game,” *Games and Economic Behavior*, 30, 163-182.
- [22] DUFWENBERG, M. AND G. KIRCHSTEIGER (2004): “A Theory of Sequential Reciprocity,” *Games and Economic Behavior*, 47, 268-298.
- [23] DUFWENBERG, M. AND M. LUNDHOLM (2001): “Social Norms and Moral Hazard,” *Economic Journal*, 111, 506-525.
- [24] ELSTER, J. (1998): “Emotions and Economic Theory,” *Journal of Economic Literature*, 36, 4774.
- [25] FALK, A. AND U. FISCHBACHER (2005): “A Theory of Reciprocity,” *Games and Economic Behavior*, forthcoming.
- [26] FEHR, E. AND S. GÄCHTER (2000): “Fairness and Retaliation: The Economics of Reciprocity,” *Journal of Economic Perspectives*, 14, 159-181.
- [27] FUDENBERG, D. AND D. LEVINE (1998): *The Theory of Learning in Games*. MIT Press.
- [28] FUDENBERG, D. AND J. TIROLE (1991a): *Game Theory*. MIT Press.
- [29] — (1991b): “Perfect Bayesian Equilibria,” *Journal of Economic Theory*, 53, 236-260.
- [30] GEANAKOPOLOS, J. (1996): “The Hangman Paradox and the Newcomb’s Paradox as Psychological Games,” Cowles Foundation Discussion Paper No. 1128.
- [31] GEANAKOPOLOS, J., D. PEARCE AND E. STACCHETTI (1989): “Psychological Games and Sequential Rationality,” *Games and Economic Behavior*, 1, 60-79.

- [32] GILBOA, I. AND D. SCHMEIDLER (1988): “Information Dependent Games: Can Common Sense Be Common Knowledge?” *Economics Letters*, 27, 215-221.
- [33] GUERRA, G. AND D.J. ZIZZO (2004): “Trust Responsiveness and Beliefs,” *Journal of Economic Behavior and Organization*, 55, 25-30.
- [34] GUL, F. AND W. PESENDORFER (2004): “The Canonical Type Space for Interdependent Preferences,” mimeo, Princeton University.
- [35] HARSANYI, J. (1967-68): “Games of Incomplete Information Played by Bayesian Players. Parts I, II, III,” *Management Science*, 14, 159-182, 320-334, 486-502.
- [36] HUANG, P.H. AND WU, H.-M. (1994): “More Order without More Law: A Theory of Social Norms and Organizational Cultures,” *Journal of Law, Economics and Organization*, 12, 390-406.
- [37] HUCK, S. AND D. KÜBLER (2000): “Social Pressure, Uncertainty, and Cooperation,” *Economics of Governance*, 1, 199-212.
- [38] KECHRIS, A. (1995): *Classical Descriptive Set Theory*. Berlin: Springer.
- [39] KOHLBERG E. AND P. RENY (1997): “Independence on Relative Probability Spaces and Consistent Assessments in Game Trees,” *Journal of Economic Theory*, 75, 280-313.
- [40] KOLPIN, V. (1992): “Equilibrium Refinements in Psychological Games,” *Games and Economic Behavior*, 4, 218-231.
- [41] KREPS, D. AND R. WILSON (1982): “Sequential Equilibrium,” *Econometrica*, 50, 863-894.
- [42] KROHNE, H.W. (2003): “Individual differences in emotional reactions and coping,” in R. J. Davidson, K. R. Scherer and H. H. Goldsmith (Eds.), *Handbook of Affective Sciences*, pp. 698-725. New York: Oxford University Press.
- [43] KUHN, H.W. (1953): “Extensive Games and the Problem of Information,” in *Contributions to the Theory of Games II*, ed. by H. W. Kuhn and A. W. Tucker. Princeton: Princeton University Press, pp. 193-216.
- [44] LEVINE, D.K. (1998): “Modeling Altruism and Spitefulness in Experiments,” *Review of Economic Dynamics*, 1, 593-622.

- [45] LOEWENSTEIN, G., E. WEBER, C. HSEE AND N. WELCH (2001): "Risk as feelings," *Psychological Bulletin*, 127, 267-289.
- [46] LI, J. (2005): "The Power of Convention: A Theory of Social Preferences," mimeo, University of Pennsylvania.
- [47] LOOMES, G. AND R. SUGDEN (1982), "Regret Theory: An Alternative Theory of Rational Choice under Uncertainty," *Economic Journal*, 92, 805-824.
- [48] MERTENS J.-F. AND S. ZAMIR (1985): "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory*, 14, 1-29.
- [49] OSBORNE, M. AND A. RUBINSTEIN (1994): *A Course in Game Theory*. MIT Press.
- [50] PEARCE, D. (1984): "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica*, 52, 1029-1050.
- [51] POPE, R. (2004): "Biases From Omitted Risk Effects in Standard Gamble Utilities," *Journal of Health Economics*, 23, 695-735.
- [52] RABIN, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83, 1281-1302.
- [53] RENY, P. (1992): "Backward Induction, Normal Form Perfection and Explicable Equilibria," *Econometrica*, 60, 626-649.
- [54] RENYI, A. (1955): "On a New Axiomatic Theory of Probability," *Acta Mathematica Academiae Scientiarum Hungaricae*, 6, 285-335.
- [55] RUFFLE, B.J. (1999): "Gift Giving with Emotions," *Journal of Economic Behavior and Organization*, 39, 399-420.
- [56] SEGAL, U. AND J. SOBEL (2003): "Tit for Tat: Foundations for Preferences for Reciprocity in Strategic Settings," mimeo, Boston College and UCSD.
- [57] STROTZ, R.H. (1956): "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, 23, 165-180.
- [58] TANGNEY, J.P. (1995): "Recent Advances in the Empirical Study of Shame and Guilt," *American Behavioral Scientist*, 38, 1132-1145.