Psychological Game Theory*

Pierpaolo Battigalli & Martin Dufwenberg

August 1, 2019

Abstract

The mathematical framework of psychological game theory is useful for describing many forms of motivation where preferences depend directly on own or others' beliefs. It allows for incorporation of emotions, reciprocity, image concerns, and self-esteem in economic analysis. We explain how and why, discussing basic theory, a variety of sentiments, experiments, and applied work.

Keywords: psychological game theory; belief-dependent motivation; reciprocity; emotions; image concerns; self-esteem

***JEL* codes:** C72; D91

1 Introduction

Economists increasingly argue that a rich variety of human motivations shape outcomes in important ways. Some categories (including profit-maximization, altruism, inequity aversion, maximin preferences, or warm glow) can be readily handled using standard tools, most notably classical game theory. However, for other important forms of motivation adequate modeling turns out to be more complicated. Consider for example the following classes:

*Battigalli: Bocconi University and IGIER, Italy; pierpaolo.battigalli@unibocconi.it. Dufwenberg: University of Arizona, USA; University of Gothenburg, Sweden; CESifo, Germany; martind@eller.arizona.edu. We have benefited from many stimulating discussions (over the years) with our coauthors of the articles cited below. For their comments and advice we thank the Editor Steven Durlauf and several referees, as well as Chiara Aina, Lina Andersson, Geir Asheim, Francesco Fabbri, Amanda Friedenberg, Kiryl Khalmetski, Rachel Mannahan, Senran Lin, Paola Moscariello, Joel Sobel, and Jin Sohn. Financial support of ERC (grant 324219) is gratefully acknowledged.

- **emotions**, including guilt, disappointment, elation, regret, joy, frustration, anger, anxiety, suspense, shame, and fear;
- **reciprocity**, or the inclination to respond to kindness with kindness and to be unkind to whoever is unkind;
- **image concerns**, e.g. when someone wants others to believe that he is smart, altruistic, or honest;
- **self-esteem**, e.g. when someone wants to believe that he is competent or brave.

These sentiments differ greatly, yet have in common that preferences depend on endogenously determined beliefs about choices and about beliefs (as we'll show!). We refer to this as *belief-dependent utility*. Standard economic models, based on classical game theory where preferences depend on choices only, are ill-equipped to model it. However, *psychological game theory* (PGT), a framework pioneered by Geanakoplos, Pearce & Stacchetti (1989) (GP&S) and further developed by Battigalli & Dufwenberg (2009) (B&D) can.[1] PGT provides a useful intellectual umbrella under which many trends in psychology and economics can be understood, related, and synthesized. The objects of analysis are called *psychological games* (p-games).

Awareness of and interest in PGT is on the rise, yet far from universal. We explain what PGT is and what motivations can be modeled, highlighting a variety of idiosyncratic features. We present old insights and speculate about new ones that PGT may hold promise to deliver. We discuss basic theory, experimental tests, and applied work. Although we cite a lot of papers, our primary goal is not to provide a comprehensive survey. Rather we try to highlight the structure and potential of various forms of work involving PGT. Our style is semi-formal, presenting some notions verbally rather than mathematically. Readers who wish to dig deeper should compare with relevant passages of GP&S, B&D, and other articles we reference. This includes the recent methodological article Battigalli, Corrao & Dufwenberg (2019) (BC&D) which contains some key innovations relative to B&D which are reflected also in our exposition here (more on that in Section 3).

[1] See also Gilboa & Schmeidler (1988) who in another pioneering contribution on "information dependent games" anticipated some of the themes that GP&S and others developed in more depth.

Our discussion of belief-dependent motivations mostly consists of showing how to represent them with psychological utility functions and highlighting the ensuing best-reply behavior. Sometimes we analyze strategic reasoning either by iterated elimination of non-best replies, or informally applying an equilibrium concept. For a broader discussion of solution concepts see B&D and BC&D, as well as our brief critical remarks in the last part of Section 7.

The sections below cover: a warm-up example suggestive of many relevant broader themes (2); the formal framework including the explanation of what a p-game is (3); how to model different forms of psychological motivations, highlighting idiosyncratic features, and mentioning some applied work (4 + 5); experimental tests (6); critical comments on methodology (7); concluding remarks (8).

2 A warm-up example

A large recent literature explores humans' reluctance to lie or cheat using an experimental "die-roll paradigm" introduced by Fischbacher & Föllmi-Heusi (2013) (F&FH).[2] Dufwenberg & Dufwenberg (2018) (D&D) propose a PGT-based account of behavior. We draw on their work for our opening example, which fits the third category of our introduction: image concerns.

A subject is asked to roll a six-sided die in private and to report the outcome, but the report is non-verifiable and can be submitted with impunity. The subject is paid in proportion to the reported number, with one exception: reporting six yields a payout of zero. We will refer to a six as a "zero." Formally, chance (player 0) draws $x \in \{0, ..., 5\}$ from a uniform distribution ($x = 0$ corresponding to rolling a six). Player 1 observes x and then chooses a report $y \in \{0, ..., 5\}$ after which he is paid y. Choice y, but not realization x, is observed by player 2, who is an "audience." In applications the audience might be a neighbor or tax authority, but in the lab could be the experimenter or an observer "imagined" by player 1. Player 2 has no (active) choice, but forms beliefs about x after observing y. The associated game-tree is G_1:

$$[G_1]$$

Numbers at end-nodes are 1's monetary payoffs, not utilities. The analysis will not depend on 2's payoffs, which are therefore not specified. The dotted

[2] See Abeler, Nosenzo & Raymond (2019) (AN&R) for a survey.

lines depict *information sets across end-nodes*. This is a feature rarely made explicit in traditional game-theoretic analysis, but here it will be critical. In the example, these sets reflect player 2's end-of-play information.

D&D explore the following preference: Player 1 feels bad to the extent that player 2 believes that 1 cheats. Measure actual cheating at end node (x, y) as $[y - x]^+ := \max\{y - x, 0\}$. Player 2 cannot observe x, but draws inferences about x conditional on y. Let $p_2(x'|y) \in [0, 1]$ be the probability 2 assigns to $x = x'$ given y, with $\sum_{x'} p_2(x'|y) = 1$, so 2's expectation of 1's cheating equals $\sum_{x'} p_2(x'|y)[y - x']^+$. Player 1's utility at (x, y) equals

$$y - \theta_1 \cdot \sum_{x'} p_2(x'|y)[y - x']^+ \tag{1}$$

where $\theta_1 \geq 0$ measures 1's sensitivity to 2's expectation of 1's cheating. Note that (1) is independent of x. This reflects that 1 cares about his image, not about cheating *per se*. Also, 1 may feel bad even if he does not lie, if the audience believes that he cheats.

We now make several points suggestive of more general PGT-themes:

First, PGT is concerned with p-games, i.e., games in which players' utility depends on endogenous beliefs. G_1 exhibits a particular instance. Player 1's utility at end node (x, y), given by (1), cannot be determined merely with reference to that end node being reached (as it would be in a standard game). Rather that utility additionally depends on 2's beliefs, via $p_2(x'|y)$.[3] Since this belief depends on 2's strategic analysis (as well as the information structure across end nodes), it is endogenous; hence, 1's preferences over outcomes are endogenous as well. The example also illustrates that simple ideas may generate a p-game. The idea that 1 feels bad to the extent that 2 believes that 1 cheats is intuitive and is easily described in words. Modeling it is straightforward, and leads to a p-game.

Second, strategic analysis of a p-game can be tractable and deliver testable predictions. Let function s_1 describe 1's *plan* (or behavior strategy)in G_1, so $s_1(x)(y)$ is the probability that s_1 assigns to y after 1 observes x. D&D solve for equilibria such that s_1 maximizes (1) given 2's beliefs, and $p_2(x'|y)$ is computed as a conditional probability using correct initial beliefs.[4] An equilibrium always exist (following B&D). However, if 1's concern for his image

[3]Formally, we need B&D's framework here, as GP&S' would not allow 1's utility to depend on *another*'s beliefs or on an *updated* belief; $p_2(x'|y)$ has both features, being 2's updated belief. We provide a more detailed comparison of GP&S and B&D in Section 3.

[4]Formally, (i) $s_1(x)(y) > 0 \Rightarrow y \in \arg\max_{y'}(y' - \theta \cdot \sum_{x'} p_2(x'|y')[y' - x']^+)$ and (ii)

is strong enough ($\theta_1 > 2$), neither honesty ($s_1(x)(x) = 1$ for all x) nor selfish choice ($s_1(x)(5) = 1$ for all x) is an equilibrium. The striking implication: if $\theta_1 > 2$ then equilibrium play involves *partial* lies (in expectation).

Walking through a sketch of the proof is helpful to get intuition for why this result holds: If honesty were expected by 2 then $p_2(x|x) = 1$ for all x, so cheating by 1 to $y = 5 > x$ would raise no suspicion, hence be 1's best response, ruling out an honest equilibrium (for any value of $\theta_1 \geq 0$) . If selfish play were expected then 2's expectation of 1's cheating would equal $\sum_x \frac{1}{6}[5 - x]^+ = 2.5$; if $\theta_1 > 2$ player 1 could then increase his utility by deviating to $y = 0$ (so that perceived cheating $= 0$).

Third, we reiterate that the analysis just conducted depends critically on the *information structure across the end nodes.* To see this, consider what would happen if those informations sets were split into singletons. That is, assume that 2 is told about both x and y, i.e. which path (x, y) occurred. At (x, y), player 2 would form beliefs such that $p_2(x|y) = 1$, implying that perceived and actual cheating coincide. If $\theta_1 > 1$ then 1's choices would be honest ($s_1(x)(x) = 1$ for all x); if $\theta_1 < 1$ then 1's choices would be selfish ($s_1(x)(5) = 1$ for all x). The partial lies prediction evaporates. This illustrates one feature (of many) that is unique to p-games. In standard games, utilities are not affected by information across end nodes, which therefore has no impact on the strategic analysis.[5]

Fourth, PGT-based predictions can be empirically relevant. The most striking insight here concerns comparing treatments that manipulate 2's information, but to get there let us first describe what F&FH found. In their data, using a design matching G_1, reporting frequencies fell between what would obtain with honest choices (16.7% for each y) and with selfish reporting (100% $y = 5$). Namely, 35% choose $y = 5$, 25% choose $y = 4$, and all other reports occur with positive frequency that declines with y. This matches D&D's partial lie prediction well (and especially an equilibrium called "sailing-to-the-ceiling"). Now couple that observation with that of the previous paragraph. If subjects were motivated as in standard game theory, then the pattern of data just described would be invariant under game variations that manipulate 2's information. Gneezy, Kajackaite & So-

$\sum_x s_1(x)(y) > 0 \Rightarrow p_2(x'|y) = \frac{s_1(x')(y)}{\sum_x s_1(x)(y)}$. Our use of equilibrium analysis in this example does not mean that we endorse it in general; see the last part of Section 7.

[5] This explains why in standard game-theoretic analysis information sets over end nodes are usually not drawn.

bel (2018) (GK&S) ran treatments where player 2 were given information about both x and y. They report that 1's behavior changed in the direction described in the previous paragraph.[6]

Fifth, having incorporated some belief-dependent motivation in a given game form, it is natural to ask whether and how that sentiment applies in other settings. The question is relevant for perceived cheating aversion as modeled by D&D, but only to a degree since that notion only makes sense in situations that permit a reporting component.[7]For other forms of motivation, one may reasonably have the ambition to extend more broadly, formulating models that apply to general classes of games. We return to that topic in Section 4,when we discuss reciprocity, guilt, anger, etc.

3 Formal framework

To ease the exposition we focus on a simple class of game forms (specifications of the rules of the game) that covers all the examples of this paper, that is, finite multistage games with monetary outcomes, in which players may move simultaneously at some stage and perfectly observe past moves (including chance moves) when they have to make a choice. However, we allow for the possibility of imperfect terminal information, which—as highlighted in the previous section—may matter for psychological reasons.[8]

The key feature of the analysis is the representation of players' beliefs about how the game is played, and their beliefs about beliefs, as such beliefs

[6]For more discussion, see Abeler et al.'s survey + meta-study + new experiments related to the F&FH's approach. They stress that "a preference for being seen as honest" is crucial for understanding the data. This covers D&D's theory and a competing approach due to GK&S and Khalmetski & Sliwka (2019) (K&S) in which a key aspect is that 1's concern with 2's opinion is based on how likely 2 believes it is that 1 lies, so $\sum_{x' \neq y} p_2(x'|y)$ rather than D&D's $\sum_{x'} p_2(x'|y)[y - x']^+$ appears in 1's utility. Also this formulation involves PGT, as again $p_2(x'|y)$ features in 1's utility.

[7]One relevant setting concerns when peers evaluate each other (e.g. in academia). Dufwenberg, Görlitz & Gravert (2019) (DG&G) extend D&D's ideas in that direction.

[8]We further simplify in two ways: First, we do not explicitly describe players' non-terminal information when they are not active, which might be relevant for some anticipatory feelings (4.2). Our analysis works "as is" under the assumption that non-active players have the coarsest information consistent with perfect recall. Second, we assume that consequences accrue at end nodes only. See BC&D for a more general and explicit analysis of time, in which the game may last for one or more periods, which may have multiple stages, and consequences accrue after each period.

affect the (psychological) utility of endnodes and expected utility calculations at non-terminal nodes. We mostly assume common knowledge of the rules of the game and of players' utility functions, i.e., complete information. Incomplete information will be addressed when we analyze specific motivations such as image concerns and self-esteem, whereby utility depends on terminal beliefs about unknown personal traits.

Our *conceptual perspective* mostly relies on B&D, rather than the seminal work of GP&S. The reason is that GP&S only encompasses utilities that depend on players' *initial* hierarchical beliefs, since at the time of their writing (i) a formal analysis of hierarchical conditional beliefs had yet to be developed, and (ii) the importance of letting utility depend on updated beliefs had not been underscored in applications. B&D instead could leverage on the recently developed theory of hierarchical conditional beliefs (Battigalli & Siniscalchi 1999) and a wealth of applications where updated beliefs enter the utility function. Motivated by conceptual arguments as well as applications, B&D substantially generalize GP&S in several ways. We will briefly point out the differences. Finally, our *formalism* relies on the recent methodological article by BC&D, which simplifies the analysis by putting only first-order beliefs of *all* players in the domain of utility (so that expected utility depends only on second-order beliefs), but sharpens other aspects, such as the representation and role of players' plans.

Game form Formally, we start with a **game form** $G = \left\langle I, \bar{H}, \iota, p_0, (\mathcal{P}_i, \pi_i)_{i \in I} \right\rangle$ with the following elements:[9]

- I is the set of **players** not including **chance**, who is player 0; the set of personal players plus chance is $I_0 = I \cup \{0\}$.
- $\bar{H}$ is a finite set of possible sequences of action profiles, or **histories** $h = \left(a^k\right)_{k=1}^{\ell}$ (for different values of ℓ, and possibly including actions of chance) with a *tree* structure: every prefix of a sequence in $\bar{H}$ (including the empty sequence $\varnothing$) belongs to $\bar{H}$ as well. Thus, histories in $\bar{H}$ correspond to nodes of the game tree and $\varnothing$ is the root. Set $\bar{H}$ is partitioned into the set of **non-terminal** histories/nodes H and **terminal** histories/nodes Z.

[9] For simplicity, we often shorten "game form" to "game."

- For each $h \in H$, $\iota(h) \subseteq I_0$ is the set of **active players**, who perfectly observe h. With this, $H_i = \{h \in H : i \in \iota(h)\}$ denotes the set of histories where i is active, and the set of feasible action profiles is

$$A(h) = \left\{(a_i)_{i\in\iota(h)} : \left(h, (a_i)_{i\in\iota(h)}\right) \in \bar{H}\right\} = \times_{i\in\iota(h)} A_i(h),$$

  where $A_i(h)$ denotes the set of feasible actions of $i \in \iota(h)$.

- p_0 is the **chance probability** function, which specifies a (discrete) probability density function $p_0(\cdot|h) \in \Delta(A_0(h))$ for each $h \in H_0$.

- For each personal player $i \in I$,

  - $\mathcal{P}_i$ is a partition of Z describing the terminal information of i that satisfies perfect recall (taking into account that active players perfectly observe non-terminal histories), $\mathcal{P}_i(z)$ denotes the cell containing z;
  - $\pi_i : Z \to \mathbb{R}$ is a material payoff function.

To illustrate, in game G_1 (Section 2), $I = \{1, 2\}$, $\bar{H} = \{\varnothing\} \cup \{0, ..., 5\} \cup Z$ with $Z = \{0, ..., 5\}^2$, $\iota(\varnothing) = \{0\}$, $p_0(x|\varnothing) = \frac{1}{6}$ and $\iota(x) = \{1\}$, $\pi_1(x, y) = y$, $\mathcal{P}_1(x, y) = \{(x, y)\}$, and $\mathcal{P}_2(x, y) = \{0, ..., 5\} \times \{y\}$ for every $(x, y) \in Z$.

Beliefs We model the **first-order beliefs** of (personal) player i as a system $\alpha_i = (\alpha_i(\cdot|h))_{h\in H\cup\mathcal{P}_i}$ of conditional probabilities about paths of play $z \in Z$. We are not assuming that i observes h when he i is *not* active at h ($h \in H\backslash H_i$). In this case we interpret $\alpha_i(\cdot|h)$ as a "virtual" conditional belief. We assume that: (0) α_i is consistent with p_0, (1) the chain rule holds, and (2) i's beliefs about simultaneous or past and unobserved actions of other players do not depend on i's chosen action.[10] The latter implies that, for each $h \in H$, i's conditional beliefs about the continuation can be obtained by multiplication from i's **plan** (behavior strategy) $\alpha_{i,i} \in \times_{h\in H_i}\Delta(A_i(h))$ and i's **conjecture** $\alpha_{i,-i} \in \times_{h\in H_{-i}}\Delta\left(A_{\iota(h)\backslash\{i\}}(h)\right)$ about co-players. Note that i's plan is part of his first-order beliefs.[11] For example, i's initially expected

[10] For example, consider a variation of G_1 where player 2 observes the report y and then bets on whether player 1 reported the truth or not. Then 2's terminal beliefs are the same as his beliefs before the bet.

[11] In the warm-up example the plan (behavior strategy) of player 1 is denoted s_1. In our abstract notation we instead write $\alpha_{i,i}$ because the plan is part of i's first-order beliefs α_i.

material payoff $\mathbb{E}[\pi_i; \alpha_i]$ (which may affect his utility *via* disappointment or frustration) depends on both $\alpha_{i,i}$ and $\alpha_{i,-i}$. As we further explain below, the interpretation is that i plans his contingent choices given his conjecture and thus ends up with an overall system of beliefs about paths.

Let Δ_i^1 denote i's space of first-order beliefs. We model **second-order beliefs** as systems $\beta_i = (\beta_i(\cdot|h))_{h \in H \cup \mathcal{P}_i}$ of conditional probabilities about both paths of play $z \in Z$ *and* co-players first-order beliefs $\alpha_{-i} \in \times_{j \in I \setminus \{i\}} \Delta_j^1$ such that: (0) the marginal beliefs about paths form a first-order belief system in Δ_i^1 (hence they are also consistent with p_0), (1) the chain rule holds, and (2) i's beliefs about α_{-i} and simultaneous or past and unobserved actions of other players do not depend on i's chosen action. We let Δ_i^2 denote the set of second-order beliefs systems of i.

To summarize, $\alpha_i \in \Delta_i^1$ denotes i's (first-order) beliefs about sequences of actions, or paths, whereas $\beta_i \in \Delta_i^2$ denotes i's (second-order) overall beliefs about paths and co-players' (first-order) beliefs. In formulas providing two-level hierarchies (α_i, β_i) we maintain the coherence assumption that α_i is the marginal of β_i.

We point out two conceptually relevant differences with B&D: (a) There, we represented behavior (what players have first-order beliefs about) as a complete description of the actions that players would take at each history where they are active, that is, a (pure) strategy profile rather than a path of play. (b) In B&D, we explicitly represented first-order beliefs as beliefs about the strategies of *others*. Our explicit interpretation in B&D was that each player knows his (pure) plan and there is a necessary coincidence between each player's plan and the objective description of how he would behave whenever active, and that such coincidence is transparent to all players (see B&D, p. 11). Here instead we follow BC&D in modeling players' beliefs about paths, hence the behavior of *everybody*.[12] Beliefs about own behavior are interpreted as (possibly non-deterministic) *plans,* which *need not coincide with actual behavior*. For example, if i is initially certain that j's plan is $\alpha_{j,j}$ and then observes a deviation from $\alpha_{j,j}$, he may still believe that j' plan was indeed $\alpha_{j,j}$ but that he took an unplanned action by mistake (a kind of "tremble" as in Selten 1975). The analysis of B&D instead rules this out: every observed action is *necessarily* interpreted as a planned choice. In sum, our framework is sufficiently expressive to model players' intentions, their

[12] The set of strategy profiles is exponentially more complex than the set of paths. Hence, beliefs about paths are simpler.

preceptions of the intentions of others, and how such perceptions are affected by observing actions. This is important in standard games to elucidate the difference between, say, forward- and backward-induction reasoning.[13] In is even more important when players care intrinsically for the intentions of others, as with many forms of belief-dependent preferences.

Standard utility Before we describe the belief-dependent utilities that are characteristic of p-games, it may be helpful to recall how utilities are defined in standard games. Namely, player i's utility has the general form $u_i : Z \to \mathbb{R}$. This does not imply that i is "selfish." Caring only about own material reward is a special case ($u_i(z) = \pi_i(z)$ for all $z \in Z$), but i could alternatively be motivated by a host of "social preferences" including altruism, inequity aversion, maximin preferences, or warm glow.[14] However, the forms of motivation we will soon exhibit are ruled out as they require a richer notion of utility.

Psychological utility and p-games As argued by BC&D, most forms of belief-dependent motivations for a given player i can be modeled by assuming that, for some terminal history z, i's utility for reaching z depends on the first-order beliefs profile $(\alpha_j)_{j \in I}$. Thus we have utility functions with the general form $u_i : Z \times \left(\times_{j \in I} \Delta_j^1\right) \to \mathbb{R}$. These typically involve both the material payoffs and some features of own or others' initial, interim, or terminal first-order beliefs. In the deception example of Section 2, player 1's utility at terminal history (x, y) depends on his monetary payoff $\pi_1(x, y) = y$ and on 2's terminal belief about die roll x given report y. In this case, utility depends on the terminal first-order beliefs of someone else. If instead i (besides liking money) dislikes disappointing j, then his utility for reaching z is decreasing j's disappointment $[\mathbb{E}[\pi_j; \alpha_j] - \pi_j(z)]^+$, which depends on j's payoff and his initial belief. In both cases, i's utility of terminal histories depends on payoffs and the (unknown) first-order beliefs of another player. This is like a standard state-dependent utility function. As noted by B&D, the maximization of its expected value can be analyzed with standard techniques leveraging on the dynamic consistency of subjective expected utility maximizers.

[13] See Battigalli & De Vito (2018) and the references therein.

[14] For prominent examples of specific functional forms, see e.g. Fehr & Schmidt (1999) (F&S), Bolton & Ockenfels (2000) (B&O), or (the main text model of) Charness & Rabin (2002) (C&R).

For other motivations like aversion to disappointment, or—more generally—expectation-based reference dependence (Kőszegi & Rabin 2006, 2007, 2009) (K&R), i's utility depends on *his* expectations (e.g., on the initially expected material payoff $\mathbb{E}[\pi_i; \alpha_i]$), hence on his own plan $\alpha_{i,i}$. We show in Section 4.2 that such forms of own-plan dependence yield dynamic inconsistency of preferences, which implies that some care is required in defining what it means to be subjectively "rational." Similar considerations apply to anticipatory feelings with negative or positive valence like anxiety, or suspense (see Section 4.2). Essentially, i's plan $\alpha_{i,i}$ must form an "intra-personal equilibrium" given his overall belief β_i. (Compare with Kőszegi 2010.) Next, we explain this in detail.

The combination of a game (form) and (psychological) utilities for all players gives a p-game. We focus mostly on p-games where the belief-dependence of utility is limited to first-order beliefs. (The exception is the part on "higher-order belief-dependence" in Section 7.)

Rational planning Fix a second-order belief $\beta_i \in \Delta_i^2$ with marginal first-order belief $\alpha_i \in \Delta_i^1$ including i's plan $\alpha_{i,i}$. For every non-terminal or terminal history $h' \in \bar{H}$, we can determine the expectation of u_i conditional on h', written $\mathbb{E}[u_i|h'; \beta_i]$. Now consider a history at which i is active, viz. $h \in H_i$. Each action $a_i \in A_i(h)$ yields expected utility

$$\bar{u}_{i,h}(a_i; \beta_i) = \sum_{a_{-i} \in \times_{j \in \iota(h) \setminus \{i\}} A_j(h)} \alpha_{i,-i}(a_{-i}|h) \, \mathbb{E}[u_i | (a_i, (a_i, a_{-i})); \beta_i].$$

Belief system β_i satisfies **rational planning** if every action that i expects to take with positive probability is a local best reply, that is,

$$\alpha_{i,i}(a_i|h) > 0 \Rightarrow a_i \in \arg \max_{a_i' \in A_i(h)} \bar{u}_{i,h}(a_i'; \beta_i)$$

for all $h \in H_i$, $a_i \in A_i(h)$. Rationality requires that i plans rationally given his beliefs and carries out his plan when given the opportunity.

When $u_i(z, \alpha)$ does not depend on α_i, or—more generally—does not depend on i's plan $\alpha_{i,i}$, then rational planning is equivalent to the standard sequential rationality condition[15] and i's rational plan can be non-deterministic (not a pure strategy) if and only if i is always indifferent between the pure

[15] The strategy of i is ex ante optimal, and the continuation strategy is optimal starting from every $h \in H_i$.

strategies in the "support" of $\alpha_{i,i}$ (cf. Remark 1 in BC&D). If instead $u_i(z, \alpha)$ depends on $\alpha_{i,i}$, first, it may be impossible to satisfy the standard sequential rationality condition, second, deterministic rational plans may not exist (see Section 4.2 for a simple example).

Local utilities and incomplete information Solution concepts for p-games can be defined and analyzed starting from the "local" utility functions $\bar{u}_{i,h} : A_i(h) \times \Delta_i^2 \to \mathbb{R}$ ($i \in I$, $h \in H_i$). To model some belief-dependent action tendencies such as the desire to reciprocate (un)kind behavior (un)kindly (Section 4.1), or the desire to vent one's own frustration by harming others (4.2), it is convenient to work directly with such history-dependent utility functions, without deriving them from utilities of terminal histories. Also, following GP&S, B&D let utility depend on beliefs of every order. In Section 7 we briefly discuss the possible relevance of k-order beliefs with $k > 2$.

A realistic analysis of strategic thinking may have to account for uncertainty about personality traits, i.e., incomplete information. This can be achieved by parameterizing such traits with some vector θ and letting players' first-order beliefs concern the unknown part of θ as well as paths of play. Beliefs about personal traits may also be essential to model some psychological motivations such as image concerns (Section 4.3) and self-esteem (4.4).

For a general analysis of the relationship between "global" and "local" utility functions and of incomplete information see BC&D and the relevant references therein.

Differences with GP&S Our perspective and formal analysis differs from that of GP&S in several ways. Let us first address the least important one: unlike GP&S (and B&D) we stop at beliefs of the second order. To our knowledge, this is enough to encompass the overwhelming majority of applications. Thus, let us focus on the case where only (first- and) second-order beliefs matter for expected utility calculations. With this, GP&S consider *initial* beliefs about the behavior and the *initial* beliefs of *others.* In particular, in games with simultaneous moves (where $Z = A := A(\varnothing)$) GP&S consider utilities of the following form: $\hat{u}_i\left(a, \beta_{i,-i}^{\varnothing}\right)$, where $\beta_{i,-i}^{\varnothing} \in \Delta\left(A_{-i} \times \left(\times_{j \neq i} \Delta\left(A_{-j}\right)\right)\right)$ denotes i's initial belief about the behavior and the (first-order) beliefs of co-players. We obtain such functional forms in the special case where only initial belief about others matter (see B&D for details). The approach of

GP&S has three important limitations. First, it rules out models where utility depends on updated beliefs, such as the warm-up example of Section 2, models of sequential reciprocity (4.1), image concerns (4.3), and self-esteem (4.4). Second, it rules out own-plan-dependent utility as in models with expectation-based reference-dependence (4.2) and anticipatory feelings (4.2). Third, even if utility depends only on initial beliefs, GP&S' framework restricts the toolbox of strategic analysis to (extensions of) traditional equilibrium concepts whereby players have correct beliefs about the (initial) beliefs of others, which therefore never change as play unfolds, on or off the equilibrium path. Indeed, if this were not the case (as in appropriate versions of rationalizability), it would be necessary to address the issue of how players update their beliefs concerning what they care about, i.e., others' beliefs.

4 Four forms of motivation

We now showcase how PGT is useful for describing many interesting forms of motivation, focusing on the classes mentioned in the introduction. Along the way we call attention to idiosyncratic technical features.

4.1 Reciprocity

The idea that people wish to be kind towards those they perceive to be kind, and unkind towards those they perceive to be unkind, is age-old and prevalent.[16] Early academic discussions can be found in anthropology (Mauss 1954), social psychology (Goranson & Berkowitz 1966), biology (Trivers 1971), and economics where the pioneer is Akerlof (1982) who analyzed "gift-exchange" in labor markets. Akerlof had the psychological intuition that reciprocity would imply a monotone wage-effort relationship, so he posited that. However, he did not engage in mathematical psychology and formal description of the underlying affective processes. Rabin (1993) realized that such an approach could bring about a generally applicable model. His is the first ever PGT-based attempt at exploring the general implications of a

[16] Fehr & Gächter (2000, p. 159) reproduce a 13^{th} century quote from the *Edda* that conveys the spirit: "A man ought to be a friend to his friend and repay gift with gift. People should meet smiles with smiles and lies with treachery." Dufwenberg, Smith & Van Essen (2013, Section III) give more examples, from popular culture, business, and experiments. Sobel (2005) provides a broad critical discussion.

particular form of motivation. For this reason, it feels natural for us to start our exploration with a look at reciprocity. Rabin focuses on simultaneous-move games but to do applied economics it is important to consider extensive games with a non-trivial dynamic structure (as Rabin pointed out; p. 1296). Dufwenberg & Kirchsteger (D&K) (2004) take on that task,[17] and we sketch their approach.

Game (form) G_2 (akin to D&K's Γ_1) is useful for introducing main ideas:

$$[G_2]$$

Define i's kindness to j – $\kappa_{ij}(\cdot)$ in D&K's notation – as the difference between the payoff i believes j gets (given i's choice) and the average of the minimum and maximum payoff j could get (for other choices of i).[18] In G_2, if 1 believes there is probability p that 2 would choose L we get

$$\begin{aligned}
\kappa_{12}(X,p) &= p\cdot 9 + (1-p)\cdot 1 - \frac{1}{2}\cdot[5 + (p\cdot 9 + (1-p)\cdot 1)] = 4\cdot p - 2\\
\kappa_{12}(Y,p) &= 5 - \frac{1}{2}\cdot[5 + (p\cdot 9 + (1-p)\cdot 1)] = 2 - 4\cdot p\\
\kappa_{21}(L) &= 1 - \frac{1}{2}\cdot[1+9] = -4\\
\kappa_{21}(R) &= 9 - \frac{1}{2}\cdot[1+9] = 4
\end{aligned}$$

Note that i's kindness to j has the dimension of the (expected, material) payoff of j, it ranges from negative to positive, and it may depend on i's beliefs (as it does for 1 in G_2). Player i is taken to maximize (the expectation of) a utility of the form

$$u_i(\cdot) = \pi_i(\cdot) + \theta_i\cdot\kappa_{ij}(\cdot)\cdot\kappa_{ji}(\cdot) \tag{2}$$

where parameter $\theta_i \geq 0$ reflects i's reciprocity sensitivity. Desire to reciprocate is captured via "sign-matching;" $\theta_i\kappa_{ij}(\cdot)\kappa_{ji}(\cdot)$ is positive only if the signs of $\kappa_{ij}(\cdot)$ and $\kappa_{ji}(\cdot)$ match.[19] To illustrate in G_2: if θ_2 is high enough, 2 wants to "surprise" 1, i.e., choose L if $p < \frac{1}{2}$ and choose R if $p > \frac{1}{2}$.

[17] The main difference between Rabin's and D&K's approaches concerns which class of games is considered, but there are other differences too. See D&K (2004, Section 5; 2019).

[18] This definition neglects an important aspect that is commented on below under the heading "Dealing with 'bombs.'"

[19] Player i cannot know j's beliefs and must consider his beliefs about κ_{ji}, called λ_{iji} by D&K who plug that second-order belief into u_i. Our formulation, (2), conformant with Section 3, relies of first-order beliefs only, but has equivalent implications to D&K's.

We make several PGT-related observations:

(i) Player 2 chooses between end nodes. In traditional games, her optimal choice would be independent of beliefs. This is not the case with reciprocity in G_2 where 2's optimal choice also depends on p, a belief of 1's. This illustrates that G_2, when players are motivated by reciprocity, leads to a p-game.

(ii) Relatedly, backward induction can not be used to find 2's optimal choice independently of beliefs. Player 2 must consult her beliefs about p.

(iii) Traditional (finite) perfect information games have equilibria (justifiable by backward induction) where players rely on degenerate plans. This is not the case in G_2, for high values of θ_2. We have not defined any equilibrium here, but suppose such a concept involves that 1 correctly anticipates 2's plan, and that 2 anticipates that. (D&K's equilibrium has that property.) If 2 plans to choose L and 1 anticipates that then $p = 1$. But if 1 anticipates that then (as explained above) 2's best response would be R, not L. An analogous argument rules out an equilibrium where 2 plans to choose R.

Our next example, the mini-ultimatum game G_3, gives further insights regarding reciprocity, and will be used for later comparisons as well:

$$[G_3]$$

Reasoning as before (with p now 1's belief about R), we see that $\kappa_{12}(G, p)$ is strictly negative for all p.[20] If θ_2 is large enough, the utility maximizing plan for 2 is R. Suppose this case is at hand. What should 1 do? If $\theta_1 = 0$, meaning that 1 is selfish, then 1 would choose F (since $5 > 0$). If instead θ_1 is large (enough), then there are two possibilities. The first one is that 1 chooses F. To get the intuition, suppose 1 believes that 2 believes (at the root) that 1 plans to choose F. Then 1 believes that 2 believes that 2 is not (as evaluated at the root) affecting 1's payoff. That is, at the root, $\kappa_{21}(\cdot) = 0$, implying that to maximize his utility 1 should act as if selfish and choose F (since $5 > 0$). The second, very different, possibility is that 1 chooses G, despite the anticipation that 2 will choose R. This is a "street fight" outcome, with negative reciprocity manifesting along the path of play. To get the intuition, suppose 1 believes that 2 believes (at the root) that 1 plans to choose G. Then 1 believes that 2 believes that 2 is generating a payoff of 0 rather than 9 for player 1. In this case, 2 would be unkind to 1. Since θ_1 is large, 1 reciprocates (in anticipation!) choosing G thereby generating a payoff of 0 rather than 5 for player 2.

[20] More precisely, $\kappa_{12}(G, p) = (1 - p) \cdot 1 - (\frac{1}{2} \cdot 5 + \frac{1}{2} \cdot [(1 - p) \cdot 1]) = -2 - \frac{p}{2}$.

The analysis here reflects a key feature of D&K's approach, namely that players' kindness is re-evaluated at each history. For example, 2's kindness to 1 at the root may be zero (if 2 believes 1 plans to choose F) and yet 2's kindness after 1 chooses G would not be zero.[21]

Dealing with "bombs" The account of reciprocity theory just given glosses over a subtle issue which we now flag for. Recall how we defined i's kindness to j as the difference between the payoff i believes j gets and the average of the minimum and maximum payoff j could get. In some games absurd implications follow unless the calculation of "the minimum payoff j could get" is modified to not consider choices that hurt both i and j. To illustrate, let G_3^X be a modification of G_3 such that player 1 has a third choice at the root – X – which explodes a bomb, leaving each player with a material payoff of -100. In G_3 we concluded that 1's kindness when choosing G was negative ($\kappa_{12}(G,p) = -2-\frac{p}{2}$, as noted in a footnote). Reasoning analogously, in G_3^X we would get that 1's kindness of choice G is instead positive.[22] Arguably, this is implausible; while hurting everyone would surely be unkind, not doing so shouldn't automatically render other choices kind. Arguably (for a given p) the kindness of choice G should be the same in G_3^X as in G_3. D&K, as well as Rabin, propose kindness definitions that achieve such an objective, by calculating "the minimum payoff j could get" without regard to so-called "inefficient strategies" that hurt both i and j.

D&K's and Rabin's approach to inefficient choice, while to a large degree similar in spirit, differ in details. The (somewhat contentious) issues involved are too subtle to warrant coverage here. We refer to D&K (2019) for a detailed discussion, including a response to a related critique by Isoni & Sugden (2019).

Related literature D&K limit attention to certain games without chance moves, a restriction Sebald (2010) drops, which allows him to address broader notions of "attribution" and "procedural concerns." Sohn & Wu (2019) (S&W) analyze games where players are uncertain about each others' reciprocity sensitivities. Jiang & Wu (2019) (J&W) discuss alternative belief-updating rules to those of D&K. Dufwenberg, Smith & Van Essen (2013) (DS&VE) modify

[21] Our account has out of necessity been sketchy; see van Damme et al. (2014; Section 6, by D&K) for a fuller analysis of a more general class of ultimatum games.

[22] More precisely, $\kappa_{12}(G,p) = (1-p)\cdot 1 - (\frac{1}{2}\cdot 5 + \frac{1}{2}\cdot(-100)) = 48.5 - p$.

D&K to focus on "vengeance;" players reciprocate negative but not positive kindness (achieved by replacing $\kappa_{ji}(\cdot)$ in (2) by $[\kappa_{ji}(\cdot)]^-$). D&K, Sebald, S&W, J&W, and DS&VE hew close to Rabin. Alternative approaches are proposed by Falk & Fischbacher (2006) who combine reciprocity motives with preferences for fair distributions,[23] and Çelen, Schotter & Blanco (2017) who model i's reciprocation to j based on how i would have behaved had he been in j's position.

As PGT-based models gain popularity they will be increasingly used to do applied economics. Most such work to date is based on reciprocity theory (and in particular D&K's model). Topics explored include wage setting, voting, framing effects, hold-up, bargaining, gift exchange, insolvency in banking, mechanism design, trade disputes, public goods, RCTs, MOUs, climate negotiations, communication, and performance-based contracts.[24]

4.2 Emotions

In a precocious article in this *Journal*, Elster (1998) argued that emotions "are triggered by beliefs" (p. 49) and that they can have important economic consequences. How "can emotions help us explain behavior for which good explanations seem to be lacking?" he asked (p. 48). He lamented economists' dearth of attention to the issue. However, PGT has subsequently been put to such use, and there is more to do. This section explains.

Guilt Psychologists Baumeister, Stillwell & Heatherton's (1994) (BS&H) argue that "the prototypical cause of guilt would be the infliction of harm, loss, or distress on a relationship partner" and that if "people feel guilt for hurting their partners ... and for failing to live up to their expectations, they will alter their behavior (to avoid guilt) in ways that seem likely to

[23] So do Charness & Rabin (2002) (C&R) in the appendix-version of their social preference model. C&R and the references in the main text are PGT-based. Levine (1998), Cox, Friedman & Sadiraj (2008), and Gul & Pesendorfer (2016) present reciprocity-related ideas which are not kindness-based and do not use PGT.

[24] See D&K 2000, Hahn (2009), Dufwenberg, Gächter & Hennig-Schmidt (2011) (DG&HS), DS&VE, van Damme et al. (2014; Section 6, by D&K), Netzer & Schmutzler (2014), Dufwenberg & Rietzke (2016), Bierbrauer & Netzer (2016), Bierbrauer, Ockenfels, Pollak & Rückert (2017) Conconi, DeRemer, Kirchsteiger, Trimarchi & Zanardi (2017), Dufwenberg & Patel (2017), Jang, Patel & Dufwenberg (2018), Kozlovskaya & Nicolò (2019), Aldashev, Kirchsteiger & Sebald (2017), Nyborg (2018), Le Quement & Patel (2018), and Livio & De Chiara (2019).

maintain and strengthen the relationship" (see p. 245; compare also Tangney 1995). That outlook is reflected in the following arguably realistic example of conscientious tipping:

> Tipper feels guilty if she lets others down. When she travels to foreign countries, and takes a cab from the airport, this influences the gratuity she gives. Tipper gives exactly what she believes the driver expects to get, to avoid the pang of guilt that would plague her if she gave less.[25]

The example can be modeled as a p-game: Let G_4 be a game form where Tipper (player 2) chooses tip $t \in \{0, 1, ..., M\}$. $M > 0$ is the amount of money in her wallet. The driver (player 1) is not active . Choice t thus pins down a strategy profile and an associated end node. Let Tipper's utility equal $t - \theta_2 \cdot [\tau - t]^+$, where τ is the driver's expectation of t. The presence of τ creates a p-game; had we had a traditional game, utilities would only be defined on endnodes, and Tipper's best choice (or choices) would be independent of τ.

Among the emotions, guilt is the one that has been explored the most using PGT.[26] B&D (2007) develop a model allowing exploration of how (two versions of) guilt shapes strategic interaction in a general class of games. While most work that connects to B&D (2007) has been experimental, a few applied theory papers explored how guilt influences marriage & divorce, corruption, cheating, framing, tax evasion, public goods, embezzlement, and expert advice.[27]

We describe B&D's (2007) notion of "simple guilt," which player i experiences when he believes that the payoff j gets ($\pi_j(\cdot)$) is lower than the payoff j initially expected ($\mathbb{E}[\pi_j; \alpha_j]$).[28] Specifically, $i \neq j$ maximizes (the

[25] When she attended an event at Bocconi, the ride from Linate was 21 Euro, and her driver said "eh, give me 20," and she was just fine with that.

[26] Reciprocity, which we do not count as an emotion, has been explored even more than guilt. See Azar (2019) for a statistical analysis of the bibliometric impact of PGT-based reciprocity and guilt theory.

[27] See Dufwenberg (2002), Balafoutas (2011), Battigalli, Charness & Dufwenberg (2013), Dufwenberg & Nordblom (2018), DG&HS, Patel & Smith (2019), Attanasi, Rimbaud & Villeval (2019) (AR&V), and Khalmetski (2019).

[28] B&D (2007) actually assume that i suffers only to the extent that he *causes* j to get a lower payoff than j initially expected. Stating that precisely, as B&D (2007) do, leads to a more complicated utility than the one seen here. However, best responses are identical, so we opt for the simpler version here.

expectation of) a utility of the form

$$u_i(\cdot) = \pi_i(\cdot) - \theta_i \cdot [\mathbb{E}[\pi_j; \alpha_j] - \pi_j(\cdot)]^+. \tag{3}$$

Again, $\theta_i \geq 0$ is a sensitivity parameter. Applied to G_4, Tipper's behavior is captured if $\theta_2 > 1$. We now discuss also a trust game G_5.[29] Assume that $\theta_1 = 0$ and $\theta_2 > 0$ to get the p-game G_5^*, displayed alongside, where q reflects 1 belief about 2's choice. Namely, 1 believes that there is probability q that 2 would choose S.[30]

$[G_5$ and $G_5^*]$

Several PGT-related observations are pertinent:

(i) G_5^* is a p-game, because of the presence of belief-feature q. One may think of 2's utility as reflecting a form of "state-dependent" preference, i.e., what 2 would prefer prefer if he knew q.

(ii) To maximize her utility, 2 must consult her beliefs about q. Early work on guilt (e.g. Dufwenberg 2002) plugged that second-order belief (rather than q) into u_2. As explained by B&D, the approaches are equivalent.[31]

(iii) If $\theta_2 > \frac{1}{q}$ then 2 prefers S over G, and vice versa. No matter how high θ_2 is, if q is low enough 2 prefers G over S. Nevertheless, 2 may reason that if 1 chose T then $q \geq \frac{1}{2}$, since otherwise 1 would not be rational. If $\theta_2 > 2$ player 2 will then prefer G over S, and if 1 believes that 2 will reason that way, he should choose T. This illustrates the potential, in some p-games, for generating powerful predictions if players reason about each others' reasoning.[32]

(iv) As argued by Charness & Dufwenberg (2006) (C&D), simple guilt can help explain why communication can foster trust & cooperation. Suppose G_5/G_5^* is augmented with a pre-play communication opportunity and that 2 *promises* 1 to choose S. If 1 believes this, and if 2 believes that 1 believes this, then simple guilt would make 2 live up to her promise. A promise by 2 feeds a self-fulfilling circle of beliefs about beliefs that S will be chosen.

[29] This is a pre-B&D (2007) setting where PGT-based guilt was discussed. See e.g. Huang & Wu (1994), Dufwenberg (2002), Dufwenberg & Gneezy (2000), Bacharach, Guerra & Zizzo (2004), and Charness & Dufwenberg (2006).

[30] Note that $\pi_2(\cdot) + \theta_2 \cdot [\mathbb{E}[\pi_1|\alpha_1] - \pi_1(\cdot)]^+ = 14 - 1 \cdot [q \cdot 10 - 0]^+ = 14 - \theta_2 \cdot q \cdot 10$.

[31] We prefer our chosen one. The shape of 2's utility is kept simpler with only first-order belief in its domain (as anticipated in Section 3 where we only considered such beliefs).

[32] Dufwenberg (2002) call this line of reasoning "psychological forward induction." See B&D and BC&D for more discussion and formalization via extensive-form rationalizability.

(v) Somewhat relatedly, in other games, one may argue that if a vulnerable party, say player i, were afraid that a guilt averse player j would take an action that could hurt i, then i would have to communicate either that he had "high expectations" or that (for given expectations) the loss of the hurtful action would be large. These ideas are, respectively, explored by Caria & Fafchamps (2019) and Cardella (2016).

(vi) B&D's (2007) model does not distinguish whether or not a belief by j is "reasonable," as regards whether or not guilt of i can be triggered. This assumption was made in order to keep things simple, and it could be unrealistic. For example, in G_4, if M is large and the driver expected Tipper to give away all she has then she might plausibly find the driver obnoxious, and enjoy giving nothing! Balafoutas & Fornwanger (2017) (B&F) and Danilov, Khalmetski & Sliwka (2019) (DK&S) discuss such "limits of guilt" (=B&F's term).

(vii) It often makes sense to assume that there is incomplete information regarding players' guilt sensitivities. See Attanasi, Battigalli & Manzoni (2016) (AB&M), Attanasi, Battigalli, Manzoni & Nagel (2019) (ABM&N), Attanasi, Battigalli & Nagel (2013) (AB&N), and Battigalli, Charness & Dufwenberg (2013, footnote 5) (BCh&D) for work that goes in that direction.

Three further observations compare simple guilt and reciprocity:

(viii) In G_5, incorporating these two sentiments imply opposite connections between q and 2's preference. Under simple guilt (i.e. in G_5^*), the higher is q the more inclined 2 will be to choose S (see (ii)). However, the higher is q the less kind is 1 (reasoning as in Section 4.1), so if 2 were motivated by reciprocity a higher q would spell less inclination to choose S.[33]

(ix) Under simple guilt, a single utility function, that depends on initial payoff expectations and on which endnode is reached, can be applied at each history where a player moves. By contrast, to capture reciprocity motivation as in D&K one must describe and re-evaluate each player's kindness at each history.[34]

[33] On this, see AB&N.

[34] Herein lies *two* differences: First, a new utility function is needed for each history; see D&K for more on this feature, which we have not illustrated very clearly since players moved but once in the games we considered. Second, since kindness depends on (foregone) choice options, game-form details matter in a way that lacks counterpart with simple guilt. See BC&D for a detailed discussion of this distinction, concerning "game-form free" vs. "game-form dependent" preferences.

(x) Some forms of belief-dependent motivation matter the most when their occurrence is counterfactual. In G_5^*, if 2 chooses S to avoid guilt, then 2 will (along the realized path) *not* experience guilt, and nevertheless guilt has shaped the outcome.[35] By contrast, if 2 were instead motivated by reciprocity, her belief-dependent motivation might be felt as she chooses S; at that time she perceives 1 as kind (in inverse proportion to q) which influences her utility as she chooses.

Disappointment Dufwenberg (2008) gives the following example which illustrates a critical role of prior expectations:

> I just failed to win a million dollars, and I am not at all disappointed, which however I clearly would be if I were playing poker and knew I would win a million dollars unless my opponent got lucky drawing to an inside straight, and then he hit his card.

Belief-dependent disappointment was first modeled by Bell (1985) and Loomes & Sugden (1986) (L&S). More recent work by K&R (and also Shalev 2000) is technically closely related, but since it is differently motivated we write about it under the separate heading of "Belief-dependent loss aversion" in section 5 below. Gill & Prowse 2011) (G&P) argue that disappointment may help explain behavior in tournaments for "promotions; bonuses; professional partnerships; elected positions; social status; and sporting trophies" (p. 495). And, because of the close connection to the belief-dependent loss aversion literature, the many topics explored there may be reinterpreted in terms of disappointment. K&R discuss consumption, risk-preferences, and savings; O'Donoghue & Sprenger (2018) (O&R) discuss other references to papers that deal with endowment effects, labor supply, job search, pricing, and mechanism design. Most of this work (not Shalev and G&P) limits attention to single decision-maker settings, but we emphasize that disappointment makes sense in games more generally.

The needed modeling machinery was in part present already in the part on guilt of Section 4.2. Factor $[\mathbb{E}[\pi_j; \alpha_j] - \pi_j(\cdot)]^+$, seen in eq. (3), captures

[35] This observation also marks a difference, to a degree, between what is the natural focus of economists and psychologists. For economists it is obvious that counterfactual experience of guilt is important, if it shapes the economic outcome. Psychologists' discussions, by contrast, tend to focus on the impact of guilt when it actually occurs. The quote with which we opened this section, from BS&H, is exceptional.

j's disappointment,[36] although in (3) it was used for the purpose of modeling i's guilt. To let i's utility reflect disappointment we can instead look at

$$u_i(\cdot) = \pi_i(\cdot) - \theta_i \cdot [\mathbb{E}[\pi_i; \alpha_i] - (\pi_i(\cdot) + k)]^+, \tag{4}$$

where $k \geq 0$. Note that $k = 0$ incorporates disappointment in the most straightforward way, while if $k > 0$ then i discounts "small" disappointments such that they have no effect on utility.[37] Below, we consider a case with $k > 0$ to make a technical point.

Utility (4) looks deceptively similar to (3) but is crucially different in that i's utility depends (in part) on i's plan. Such "own-plan dependence," where i's beliefs about his choices impacts the utility of his choices, can lead to subtle complications as we now highlight (and see BC&D for more).

While (4) is applicable to any game form, and hence can shape strategic interaction generally, the clearest way to exhibit the essence of disappointment is to use a one-player game. We will work with game form G_6. Assume that $0 < x < 1$ while $0 \leq k \leq \min\{x, 1 - x\}$.

$$[G_6]$$

Can *Stay* be an optimal plan for 1 in G_6 (with utility given by (4))? This requires

$$\underbrace{x}_{\substack{\text{utility of } \textit{Stay} \\ \text{after planning } \textit{Stay}}} \geq \underbrace{\frac{1}{2} \cdot 2 - \frac{1}{2} \cdot \theta_1 \cdot [x - (0 + k)]^+}_{\text{utility of } \textit{Bet} \text{ after planning } \textit{Stay}} \iff x \geq \frac{2 + \theta_1 \cdot k}{2 + \theta_1}. \tag{5}$$

Similarly, *Bet* is an optimal plan if

$$\underbrace{\frac{1}{2} \cdot 2 - \frac{1}{2} \cdot \theta_1 \cdot [1 - (0 + k)]^+}_{\text{utility of } \textit{Bet} \text{ after planning } \textit{Bet}} \geq \underbrace{x - \theta_1 \cdot [1 - (x + k)]^+}_{\text{utility of } \textit{Stay} \text{ after planning } \textit{Bet}} \iff x \leq \frac{2 + \theta_1 - \theta_1 \cdot k}{2 + 2 \cdot \theta_1}. \tag{6}$$

[36] This suggests an alternative way to think of i's guilt towards j, namely that i is averse to j being disappointed.

[37] Disappointment aversion may violate first-order stochastic dominance (FSD). For example, if k in eq. (4) is 0 and $\theta_i > 1$, then i prefers a sure payoff $x > 0$ to the lottery that yields x and $2x$ with 50% chance. The axiomatization of Gul (1991) rules this out. Cerreia-Vioglio, Dillenberger & Ortoleva (2018) derive an explicit representation of preferences à la Gul (1991).

(i) First, assume that $k = 0$. Inspecting (5) and (6) one sees that if $x \in [\frac{2}{2+\theta_1}, \frac{2+\theta_1}{2+2\theta_1}]$ then either *Stay* or *Bet* can be an optimal plan. If $x \in (\frac{2}{2+\theta_1}, \frac{2+\theta_1}{2+2\theta_1})$ then 1 incurs a loss if he deviates from the plan. Such multiplicity of strictly optimal plans could never happen without own-plan dependent utility.[38] In the standard case, multiplicity of optimal plans is possible only if there is indifference.

(ii) An interesting variation arises if $k > 0$. Could it be that neither *Stay* nor *Bet* is an optimal plan? If so, then neither (5) nor (6) would hold, and we would get

$$\frac{2 + \theta_1 \cdot k}{2 + \theta_1} > x > \frac{2 + \theta_1 - \theta_1 \cdot k}{2 + 2 \cdot \theta_1}. \tag{7}$$

To see that this is possible, pick a case that is easy to compute: assume that $x = k = \frac{1}{2}$, and study (7) as θ_1 increases. The leftmost term exceeds $\frac{1}{2}$ for any $\theta_1 \geq 0$, while the rightmost term is lower than $\frac{1}{2}$ for high enough θ_1 (it decreases from 1 to $\frac{1}{4}$ as θ_1 goes from 0 to infinity). All in all, for a high enough value of θ_1, (7) must hold.

(iii) The emotion of *elation*, discussed by Bell and L&S, is a sort of opposite of disappointment. It can be modeled by substituting $[\cdot]^-$ for $[\cdot]^+$ in (4) which then leads to p-games.

Frustration Psychologists argue that people get frustrated when they are unexpectedly denied things they care about. That sounds like disappointment! However, while disappointment is mainly discussed in regards to pangs incurred and anticipated, frustration is more often discussed for how it influences decision making going forward. In particular, there is the "frustration-aggression hypothesis," originally proposed by Dollard et al (1939) (see also e.g. Averill 1982, Berkowitz 1978, 1989, Potegal, Spielberger & Stemmler 2010), whereby frustration breeds aggression towards others. We limit our discussion of frustration to its role in that context, which, we argue, suggests a difference in how to model frustration and disappointment. This will be discussed next, under the heading of ...

[38] This statement is true if there is perfect recall; otherwise similar complications occur as again dynamically inconsistent preferences may appear, and the conditional expected utility of actions may depend on the planned probability of choosing "earlier" actions. See, e.g., Piccione & Rubinstein (1997), which is the lead article in a special issue devoted to imperfect recall.

Anger Frustration breeds anger and aggression toward others. This can have profound economic impact, though few economists studied the topic. Battigalli, Dufwenberg & Smith (2019) (BD&S) propose a broadly applicable model. They do not develop applications, but mention pricing, domestic violence, riots, recessions, contracting, arbitration, terrorism, road rage, support for populist politicians, and bank bail-outs as potentially interesting ones.[39] We sketch key features of BD&S's approach, starting with an example of theirs – G_7 – designed to make a technical point about frustration:

$$[G_7]$$

Suppose that if 2 is frustrated she will consider 1 an attractive target of aggression. What would she do if 1 chooses F? The answer may seem intuitively obvious, but consider what would happen if frustration were modeled as disappointment (more disappointment giving higher inclination to aggression). Building on eq. (4), there would be multiple optimal plans for 2, following the logic of (ii) in the disappointment part of Section 4.2. If 2 plans to choose d, and if she believes 1 will choose D, then she would be disappointed after F, hence choose d in order to hurt 1.

With outcome $(2, 2)$ available, this seems psychologically implausible. BD&S instead assume that when a player evaluates her frustration, she focuses on what happened and what she can best achieve, going forwards, in material terms. Maybe she will be frustrated and end up meting out a costly punishment, but that should be a reaction to rather than a cause of her frustration. This consideration leads BD&S to the following definition of i's frustration at history h:

$$\mathrm{F}_i(h;\alpha_i) = \left[\mathbb{E}[\pi_i;\alpha_i] - \max_{a_i \in A_i(h)} \mathbb{E}[\pi_i|(h,a_i);\alpha_i]\right]^+. \tag{8}$$

Applied to G_7, let p be the probability 2 initially assigns to F while q is the probability with which 2 plans to choose f. We get $\mathrm{F}_2(F;\alpha_i) = [(1-p)\cdot 1 + p\cdot q\cdot 2 - 2]^+ = 0$. Zero frustration breeds no aggression, so 2 will choose f.

While eq. (8) differs from the disappointment-part of (4), it is still a feature that brings own-plan dependence and belief-dependent p-game utilities.

[39] As BD&S discuss, some of these topics have been analyzed by other authors empirically or using models that feature anger which however is not modeled using PGT. See e.g. Rotemberg (2005, 2011) on pricing, Card & Dahl (2011) on family violence, and Passarelli & Tabellini (2017) on political unrest.

. Having defined frustration, the next step is to model how that breeds anger and frustration. We avoid going into technical details – see BD&S for that – and here just highlight some key themes. Number one is that one must now theorize about blame. Consider G_8 (where players payoffs are listed in alphabetical order, and Abe is a dummy player):

$$[G_8]$$

BD&S assume that a frustrated player (which in G_8 could only be Penny because (8) must equal 0 at the root) becomes inclined to hurt those deemed blameworthy. They develop three models based on different blame notions:

(i) *Simple anger*: all co-players are blamed independently of how they have chosen.[40] In G_8, if Penny's anger sensitivity θ_{P} is high enough, she would choose A, going after Don whom she is most efficient at punishing.

(ii) *Anger from blaming behavior*: i's co-players are blamed to the extent that they could have averted i's frustration had they chosen differently. In G_8, with θ_{P} high, Penny would choose B, going after Bob, since Don is no longer blameworthy (he had no choice!), and Penny is more efficient at beating up Bob than Abe.

(iii) *Anger from blaming intentions*: i's co-players are blamed to the extent that i believes they intended to cause i's frustration. In G_8, with θ_{P} high, Penny would choose A, going after Abe, since also Ben is no longer blameworthy (while he could have averted Penny's dismay, he had no rational way of correctly figuring out chance's choice, and thus can't have had bad intentions). This third category, because Penny cares about others' intentions, injects a second form of belief-dependence in players' utilities.

Finally, a comment about how BD&S's models apply to the mini-ultimatum game, G_3. A comparison with reciprocity theory is of interest, since both approaches can help explain the prevalence of fair offers (F) and rejections (R). In both cases (D&K and BD&S) 2 may rationally plan to choose R (if θ_2 is high enough, and, in the case of BD&S, if 2's initial belief that 1 will choose F is strong enough). However, whereas in D&K's theory it is possible that 1 chooses the greedy offer G even if he expects 2 to choose R (since 1 then views 2 as unkind, and so may want to retaliate), this could never happen

[40] Some psychologists argue that frustrated people tend to be unsophisticated and inclined to blame in such a way. It seems to us that how and why people blame is an interesting empirical issue, which may depend on e.g. how tired a person is or on whether he or she has drunk a lot of beer.

in (any of the three versions of) BD&S' theory. As hinted at in the previous paragraph, at the root a player cannot be frustrated and he must therefore maximize his expected material payoff.

Valence and action-tendency This is a good time to insert a general reflection about emotions and p-games: Emotions have many characteristics, two important ones being *valence*, meaning the costs or rewards associated with an emotion, and *action-tendency*, or how an emotion's occurrence incites new behavior. When modeling emotions using PGT one needs to choose which aspect to highlight, or abstract from. For example, B&D's (2007) models of guilt are all about valence, abstracting away from action-tendency (which could be restrictive; see e.g. Silfver 2007 on "repair behavior"). By contrast, BD&S's models of anger are all about action tendency, as frustration has no valence in their models (again, a restrictive abstraction, as frustration may have similar valence as disappointment).

Regret Despite Édith Piaf's assertion, regret can be a powerful feeling. Bell (1982) and Loomes & Sugden (1982) (L&S) develop theories, focusing on pairwise choice, and Quiggin (1994) proposes an extension for general choice sets. These authors restrict attention to single decision maker settings, but regret makes equal sense with strategic interaction. Economists have not discussed regret much, however, the main exception being some papers on regret in auctions.[41] To consider other settings, however, one needs a general model. B&D, BC&D, and Dufwenberg & Lin (2019) formulate relevant definitions. We explain why (unlike in the case with disappointment) PGT is not needed for handling the decision theorists' settings, and why nevertheless PGT is crucial for analyzing games.

Consider the following version of Quiggin's approach: Let Ω and C be (finite) sets of states and consequences. Let $A \subseteq C^{\Omega}$ be the non-empty set of feasible acts. A decision maker ($= 1$) chooses an act, and $v_1 : C \to \mathbb{R}$ describe 1's "choiceless utility" (L&S' terminology) of consequences. However, after 1 chooses $a \in A$ chance's choice $\omega \in \Omega$ is revealed and 1 now ruminates on what-could-have-been. His regret-adjusted utility, which is what he wants to

[41] See e.g. Engelbrecht-Wiggans (1989), Engelbrecht-Wiggans & Katok (2008), Filiz-Ozbay & Ozbay (2007). These studies are not PGT-based.

maximize, is a function $u_1 : \Omega \times A \to \mathbb{R}$ defined by

$$u_1(\omega, a) = v_1(a(\omega)) - f(\max_{a' \in A} v_1(a'(\omega)) - v_1(a(\omega))), \tag{9}$$

where $f : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing. For our purposes it is useful to re-formulate this as a one-player game with a chance-move, with perfect information at end nodes: Chance makes a choice from Ω. Player 1 is not informed of chance's choice, and chooses $a \in A$. Then endnode (ω, a) is reached and revealed to 1, whose utility is computed using (9). Note that this is a standard game, because 1's utility is uniquely defined at each endnode.

However, if one generalizes the above steps to apply to any game form, then one arrives at a p-game: To see this, fix an extensive game form, focus on player i, and try to compute his regret-adjusted utility at end node z (and associated information set). To do that, one needs to know which choices i's co-players actually made, and which ones they would make at any history in the game tree that i could have made play reach had he chosen differently than he did. And that computation, of course, will reach a different answer dependent on which choices the co-players are assumed to make. In contrast to the single-player example of the previous paragraph, i's regret-adjusted utility will not be uniquely defined. If i regret-adjusts based on his beliefs about what would have happened had he chosen differently, we get a p-game. The belief-dependence of i's utility involves his own beliefs at end nodes (and associated information sets) regarding co-players' choices.

For example, consider G_2. Would 1 experience regret if he chose X, and if so how much? The answer depends on p, the probability with which 1 believes that 2 would choose L had 1 chosen Y. Analogous remarks apply to e.g. G_3, G_5, and G_7.

Anticipatory feelings So far we considered either the action tendencies caused by emotions, as in the frustration/aggression hypothesis, or how actions cause emotions (own or of others) with positive or negative valence and how players take this into account in their choice, as in guilt, disappointment or regret aversion. Behavior of the second kind is explained by the anticipation of future feelings under different courses of action. Now we consider how uncertainty about the future can cause "anticipatory feelings" with negative or positive valence in the present (cf. Loewenstein, Hsee, Weber, & Welch 2001). Of course, the anticipation of such anticipatory feelings can drive behavior in earlier periods. Timing is essential to model anticipatory feelings.

The simplest setting for a meaningful discussion is one with three dates—0, 1, 2—comprising two periods t between dates $t-1$ and $t \in \{1,2\}$. Action profile a^t is selected in period t. To make the problem interesting, player i—the decision maker under consideration—has to be active in period 1 and another player (typically, chance) has to be active in period 2.

Anxiety is an anticipatory feeling with negative valence caused by uncertainty about future material outcomes (e.g., health, or consumption). Drawing on earlier work by Kreps & Porteus (1978) on preferences for the temporal resolution of uncertainty, Caplin & Leahy (2001) put forward an axiomatic model of utility of "temporal lotteries" and consider specific functional forms. As one example, they analyze portfolio choice. Using our notation, they consider the following utility

$$u_i\left((a^1,a^2),\alpha\right) = -\left(\theta_i^V \mathbb{V}\left[\pi_i|a^1;\alpha_i\right] - \theta_i^E \mathbb{E}\left[\pi_i|a^1;\alpha_i\right]\right) + v_i^2\left(\pi_i\left(a^1,a^2\right)\right), \tag{10}$$

where $\mathbb{V}$ is the variance operator, $\theta_i^V, \theta_i^E \geq 0$, and v_i^2 is the period-2 utility of the realized material outcome. Their theory helps explain the risk-free rate puzzle and the equity-premium puzzle: when buying safe assets an agent is "paying for his peace of mind."

Caplin & Leahy also briefly mention how their general theory can be adapted to model suspense, i.e., the pleasure experienced immediately prior to the anticipated resolution of uncertainty. This theme is explored in depth by Ely, Frankel, & Kamenica (2015). Finally, Caplin & Leahy (2004) draw on their (2001) theory to study interaction between, e.g., an anxious patient and his caring doctor, who decides whether or not to reveal information affecting the patient's anticipatory feelings.

Elster's list While we have covered several emotions, and highlighted their connections with PGT, we have not considered Elster's (1998) full list. He discussed 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. We suspect that many of the additional sentiments involve belief-dependent motivation that could be explored using PGT. However, rather than pursue these topics we propose that they hold promise for rewarding research to come.

4.3 Image concerns

Introspection, empirical, and experimental evidence suggest that people are willing to give up some material payoffs to improve the opinion of others about them. We explained in Section 2 how the experimental results of F&FH about deception can be explained by a trade-off between monetary payoff and a reduction of the perceived extent of cheating or lying (D&D, GK&S, K&S). Other models instead assume that agents try to signal that they have "good traits," e.g., that they are altruistic or fair (e.g., Bénabou & Tirole 2006; Andreoni & Bernheim 2009; Ellingsen & Johannesson 2008; Grossman & van der Weele 2017), which may explain behavior in the Dictator Game, or why people seldom give anonymously to charities, while they are happy to give non-anonymously (as shown by Glazer & Konrad 1996). Several other articles explore various forms of image concerns explaining, e.g., conformity, job-seeking effort, randomized survey-response, shame avoidance, peer evaluations, and pricing distortions.[42]

The aforementioned examples suggest two broad kinds of image about which people are concerned: others' (terminal) beliefs about (a) imperfectly observed *bad/good actions*, and (b) imperfectly observed *bad/good traits.* Both are modeled by psychological utility functions.

Opinions about bad/good actions Suppose for simplicity that, according to some standard, paths in Z_i^B (resp. Z_i^G) are such that player i behaved in a bad (resp. good) way. Some paths may be neutral, e.g., because i did not play. For example, in the cheating game of Section 2 $Z_i^B = \{(x, y) : y \neq x\}$ could be the set of paths where i lies, and $Z_i^G = Z \backslash Z_i^B$. In a Trust Minigame (e.g., G_5 above and G_9 below) i is the trustee and Z_i^B contains (resp. Z_i^G) the paths where he grabs (resp. shares).[43] Let j be an observer and let $p_{j,i}^B(z) = \alpha_j\left(Z_i^B|\mathcal{P}_j(z)\right)$ (resp. $p_{j,i}^G(z) = \alpha_j\left(Z_i^B|\mathcal{P}_j(z)\right)$) denote the observer's ex post probability of bad (resp. good) deeds. An image concern related to bad/good deeds can be captured by a simple functional form like

$$u_i\left(z, \alpha\right) = \pi_i\left(z\right) + \theta_i\left[p_{j,i}^G(z) - p_{j,i}^B(z)\right]. \tag{11}$$

[42] See Bernheim (1994), Dufwenberg & Lundholm (2001), Blume, Lai & Lim (2019), Tadelis (2011), DG&G, and Sebald & Vikander (2019). We note that some of the cited models of image concern do not make the PGT-connection explicit.

[43] Note that paths record the behavior of every active player, hence we can accomodate norms such as behaving (or not) like the majority.

More generally, one can assume that i cares about the perceived distance from the standard rather than mere compliance (as in D&D), or that intrinsic motivations—besides image concerns—also play a role (i (dis)likes good (bad) deeds as in GK&S and K&S).

Opinions about bad/good traits The second kind of image concern starts from intrinsic motivation. People have heterogeneous intrinsic motivations to do good deeds and avoid bad ones, and are imperfectly informed about the motivations of others. This expands the domain of uncertainty. Thus, we have to assume that each player j has a system of conditional beliefs α_j about paths *and* traits of others.[44]In particular, terminal beliefs of j have the form $\alpha_j\left(\cdot|\mathcal{P}_j\left(z\right)\right) \in \Delta\left(\mathcal{P}_j\left(z\right) \times \Theta_{-j}\right)$. Let $\mathbf{I}_i^D\left(\cdot\right) : Z \to \{0,1\}$ denote the indicator function of Z_i^D (bad/good deeds, $D = B, G$). Intrinsic motivation is measured by parameter $\theta_i^{\mathbf{I}} \geq 0$, and player i—besides liking material payoff and being intrinsically motivated—also cares about j's estimate of $\theta_i^{\mathbf{I}}$ as (for example) in the additive utility function

$$u_i\left(z, \alpha, \theta_i\right) = \pi_i\left(z\right) + \theta_i^{\mathbf{I}}\left[\mathbf{I}_i^G(z) - \mathbf{I}_i^B(z)\right] + \theta_i^{\mathbf{O}}\mathbb{E}\left[\widetilde{\theta_i^{\mathbf{I}}}|\mathcal{P}_j\left(z\right); \alpha_j\right]. \tag{12}$$

Utility functions like (12) introduce a familiar element of signaling into the strategic analysis: even if i's intrinsic motivation to do good ($\theta_i^{\mathbf{I}}$) is low, he may be willing to pay a material cost to make j believe that $\theta_i^{\mathbf{I}}$ is high, hence that i is a "good guy." The simplest models of this kind are signaling games where only the sender is active and the receiver is a mere observer.[45]

A noteworthy application of this approach concerns privacy. Depending on available technology and regulation, what we do may be monitored even when it does not affect the material payoff of anybody else. Many people seem to care about this and in Western countries there is a consensus that privacy should be protected. Gradwohl & Smorodinsky (2017) model this by considering functional forms such that (as in eq. 12), for each (z, θ_i), $u_i\left(z, \alpha, \theta_i\right)$ depends on the terminal belief $\text{marg}_{\Theta_i}\alpha_j\left(\cdot|\mathcal{P}_j\left(z\right)\right)$ of the "audience" about θ_i. In particular they assume that—other things being equal—the agent either

[44]Formally, we are considering beliefs in games with incomplete information; see Section 5 below (under the heading "Private sensitivities") for a discussion.

[45]The signaling element may also be present in p-games with utility like (11): in the warm-up example of Section 2 report y is a(n imperfect) signal about the die roll x.

wants j's posterior to be the same as the prior,[46] or dislike being identified.[47] Focusing on the simple case where i is the only active agent and actions are observable by j (lack of privacy), they analyze the pooling and separating (Bayesian perfect) equilibria of the resulting signaling game. Pooling distorts actions from the first best that would obtain under perfect privacy. Separation inflicts a psychological utility loss due to identification.

4.4 Self-esteem

Self-esteem reflects an individual's overall subjective emotional evaluation of his own worth. It is "the positive or negative evaluations of the self, as in how we feel about it" (Smith & Mackie, 2007). We can model self-esteem by assuming that a valuable personal trait $\theta_{0,i}$ of player i chosen by nature (same index as chance) is imperfectly known by i. Such trait could be general intelligence, or ability. Player i's utility is increasing in his estimate of $\theta_{0,i}$, as in function

$$u_i\left(z,\alpha,\theta\right)=\pi_i\left(z,\theta\right)+v_i^e\left(\mathbb{E}\left[\widetilde{\theta_{0,i}}|\mathcal{P}_i\left(z\right);\alpha_i\right]\right), \tag{14}$$

where the "ego-utility" v_i^e is increasing, and we allow π_i to depend on parameter vector θ because traits such as ability typically affect material outcomes. For example, Mannahan (2019) shows that if π_i is observed ex post and v_i^e is concave, i may decide to handicap himself ensuring a bad outcome (e.g., by not sleeping before an exam) rather exposing himself to the risk of discovering that his ability is low.[48]

Also, better informed players may engage in signaling to affect i's self-esteem: Does a teacher want to reveal to a student how bad his performance was? Better information may allow for a better allocation of the student's time (more study, less leisure), but it may also be detrimental: by decreasing the student's estimate of his ability it can bring it in a range where ego-utility is more concave and cause the self-handicapping effect described above.

[46] That is, $\alpha_{j,\Theta_i}^0 \in \arg\max_{\mu_j\in\Delta(\Theta_i)} u_i\left(z,\mu,\theta_i\right)$ for each (z,θ_i), where $\alpha_{j,\Theta_i}^0 = \text{marg}_{\Theta_i}\alpha_j\left(\cdot|\varnothing\right)$ is the exogenous and known prior of j about θ_i.

[47] That is, $\delta_{\theta_i}\in\arg\max_{\mu_j\in\Delta(\Theta_i)} u_i\left(z,\mu_j,\theta_i\right)$ for each (z,θ_i), where δ_{θ_i} is the degenerate (Dirac) marginal belief that assigns probability 1 to θ_i.

[48] There is a discussion in psychology of similar self-handicapping strategies, with implications regarding for example drug use. Berglas & Jones (1978) is a classic experimental study on this topic.

A few economic studies of self-esteem model utility in line with our description here, although (unlike Mannahan) they do not make the PGT-connection explicit. See Kőszegi (2006), Eil & Rao (2011), Mobius, Niederle, Neihaus & Rosenblat (2011), and Kőszegi, Loewenstein & Murooka (2019).

5 More motivations, and related issues

Section 4 was structured around the four categories of motivation mentioned in the introduction. We now complement that presentation with sundry related themes: additional forms of belief-dependent motivation as well as some aspects of broader relevance.

Opposites Sometimes a meaningful belief-dependent motivation takes an "opposite" form of another sentiment. We already saw an example in Section 4.2 where elation was compared to disappointment (see (iii) in that part) and later on (under "Anticipatory feelings") suspense was related to anxiety. Another example involves an opposite to guilt. Ruffle (1997) and Khalmetski, Ockenfels & Werner (2015) (KO&W) consider that i enjoys surprising j so that j gets a higher material payoff than j expected. See also Dhami, Wei & al-Nowaihi (2019) (DW&a-N). This can be modeled by substituting $[\cdot]^-$ for $[\cdot]^+$ in (4).

We did not give elation or joy-of-surprising their own heading in Section 4.2, for different reasons. Elation is not discussed nearly as often as disappointment, and seems to be less often regarded as empirically relevant.[49] As regards enjoying surprising others, is that an "emotion"? Maybe yes; obviously it is a kind of joy, and joy is often listed as an emotion. However, the sentiment feels rather special and since we feel that we do not have much to say about other forms of joy we elected not to develop that category further, leaving it for future research.

As a slight aside, we also note that desire to surprise has venerable PGT-ancestry. GP&S explored the idea in their verbally presented opening example, although a different variety than the work cited above. GP&S's example does not require that the co-player is surprised in terms of material pay-

[49] In line with that, G&P report results indicating "that winners are elated while losers are disappointed, and that disappointment is the stronger emotion" (p. 495).

off.[50] Here is the quote (from p. 62), illustrating the sentiment and a feature idiosyncratic to p-games:

> Think of a two-person game in which only player 1 moves. Player 1 has two options: she can send player 2 flowers, or she can send chocolates. She knows that 2 likes either gift, but she enjoys surprising him. Consequently, if she thinks player 2 is expecting flowers (or that he thinks flowers more likely than chocolates), she sends chocolates, and vice versa. No equilibrium in pure strategies exists. In the unique mixed strategy equilibrium, player 1 sends each gift with equal probability. Note that in a traditional finite game with only one active player, there is always a pure strategy Nash equilibrium. That this is untrue in psychological games demonstrates the impossibility of analyzing such situations merely by modifying the payoffs associated with various outcomes: any modification will yield a game with at least one pure strategy equilibrium.

Higher-order belief-dependence The framework presented in Section 3 restricts the domain of a player's utility to depend on beliefs (own and others') up to only the first order.[51] This is enough to handle almost all forms of motivation that to date have been modeled using PGT.[52] The main exception is B&D's (2007) model of guilt-from-blame (but see also B&D 2009, p. 14). We now sketch how that sentiment works in an example designed to provide a contrast with simple guilt (as presented in Section 4.2). Guilt-from-blame plugs a third-order belief into the domain of a player's utility, so we leave the framework of Section 3. Our account will mainly be verbal and intuitive:

> First, for each end node z in a game, measure how disappointed j is as $[\mathbb{E}[\pi_j; \alpha_j] - \pi_j(z)]^+$ (compare (3) & (4)). Second, calculate

[50] Yet another example appears in Geanakoplos (1996), which reconsiders the classical "hangman's paradox" from philosophy, where the desire to surprise has a sadistic flavor.

[51] Player i may still have to consider his second-order beliefs, if his utility depends on j's first-order beliefs (as it did in our presentation of reciprocity, guilt, anger from blaming intentions, and image concerns). Since i does not know j's beliefs, he has to form beliefs about them to calcuate a best response.

[52] This includes reciprocity, if formulated as in Section 4.1. (As we noted in a footnote there, Rabin, D&K, and others use a different formulation with utilities that depend on second-order beliefs.)

how much of $[\mathbb{E}[\pi_j;\alpha_j]-\pi_j(z)]^+$ could have been averted had i chosen differently. Third, calculate i's initial belief regarding $[\mathbb{E}[\pi_j;\alpha_j]-\pi_j(z)]^+$. Fourth, for each z, calculate j's belief regarding i's initial belief regarding $[\mathbb{E}[\pi_j;\alpha_j]-\pi_j(z)]^+$; this is how much j would blame i if j knew he were at z. Finally, i suffers guilt-from-blame in proportion to j's blame, and i's utility trades off avoidance of that pang against i's material payoff.

B&D (2007; see Observation 1) prove that simple guilt and guilt-from-blame sometimes have comparable implications. However, this is not true in general. To illustrate, consider G_9, a modified version of G_5 in which even if 2 chooses S there is a $\frac{1}{6}$ probability that 1 gets a material payoff of 0. Moreover, if 1 gets 0 then 1 is not informed of 2's choice. As in Section 4.2, 1 believes that there is probability q that 2 would choose S.

$$[G_9]$$

Everything we said about simple guilt and (3) in Section 4.2 we could have said as regards G_9 rather than G_5. We used G_5 merely because it is more spare, but C&D (cited under (iv) in the guilt part of Section 4.2) actually used G_9 rather than G_5.[53]

If player 2 is instead motivated by guilt-from-blame then the implications are very different in G_5 and G_9. In G_5, following *In*, if player 2's second-order beliefs assign probability 1 to $q=1$, then for a high enough θ_2 player 2's best response is S. This is true just as it would be also under simple guilt. In G_9, however, following *In*, if player 2's second-order beliefs assign probability 1 to $q=1$, then player 2's best response is G regardless of how high θ_2 is! To appreciate why, note that if 2 believes that $q=1$ then 2 believes that 1 will not blame 2 if 2 chooses G, so 2 can do this with impunity.[54]

Game G_9 with guilt-from-blame joins our warm-up example (Section 2) in illustrating the critical role information across endnodes can play in p-games. Modify G_9 such that 2's doubleton information set is broken up into two singletons. That is, if 1 gets 0 then 1 *is* informed of 2's choice.[55] The

[53] C&D's reason is conceptual; from a contract-theoretic point G_9 may be seen to incorporate an element of "moral hazard" which is absent in G_5. See C&D (p. 1582).

[54] The logic here is similar to that we illustrated under the "second" point made in regards to the warm-up example in Section 2.

[55] Tadelis compares behavior in experimental treatments that resemble G_9 as well as the variation that we are describing here.

logic of the previous paragraph no longer applies, and in the modified version of G_9 guilt-from-blame and simple guilt again work similarly.

Belief-dependent loss aversion When we discussed disappointment, in Section 4.2, we mentioned how that sentiments is closely related to ideas explored by K&R and by Shalev. The goal of these authors, however, is not to model disappointment, but rather to tie in with Kahneman & Tversky's (1979) (K&T) work on prospect theory. K&R model K&T's central notion of a "reference level" as a decision maker's initially expected outcome. When he gets less than he expects he experiences loss, effectively much like in disappointment theory. K&R allow for losses in many dimensions, e.g. in $n + 1$ dimension if there are n goods as well as money. To capture that we would have to augment the framework in Section 3, to allow π_i to be vector-valued.

The key features we highlighted in regards to disappointment in Section 4.2 have counterparts in the work of K&R. Most notably the feature of own-plan dependent utility is there, and it may lead to multiplicity of optimal plans, as well as non-existence of degenerate rational plans. Beyond those technical similarities, details differ quite a lot. The exact way in which K&R define belief-dependent loss is different from the way that Bell and L&S (and we) define disappointment, and K&R also consider more notions of rational planning than we did when we discussed disappointment. The recent and penetrating survey on "Reference-Dependent Preferences" by O&R discusses all of these aspect in depth, so we refer to their text (and especially their Sections 5-7) for further details.

Social norms Fehr & Schurtenberger (2018) define a social norm as a commonly known standard of behavior that is based on widely shared views of how individual group members ought to behave in a given situation. Similar ideas are discussed by Elster (1989), Bicchieri (2005), Andrighetto, Grieco & Tummolini (2015) (AG&T), and Cartwright. D'Adda, Dufwenberg, Passarelli & Tabellini (2019) (d'ADP&T) develop a model for a restrictive context (a form of dictator games) where the central notions concern a player's conception of "the right thing to do" and a proclivity to do what *others* think is the right thing to do, especially if there is consensus about this (which would then be a social norm). The following quote from d'ADP&T reflects how this exercise is related to PGT:

> Departing from a social norm entails an element of disappointing the expectations of others, and we explore the idea that decision makers are averse to doing so. In this regard, the motivation we look at resembles guilt aversion (see B&D 2007 for a general model), a belief-dependent sentiment the modeling of which requires the framework of PGT (GP&S; B&D 2009). However, we consider expectations regarding how one "ought to behave", not how one will actually behave, which marks a way that our approach is not formally captured by p-games as formulated in the papers we cited.

Many scholars have written papers about social norms, but few proposed formal models, in particular models that can be generally applied.[56] There is work to do in this arena, and we suggest that it should involve (some possibly extended version of) PGT.[57]

Emotion carriers In most of the models we discussed above, the belief-dependent part of a player's utility was built up with reference to particular material payoffs. For example, following B&D (2007), player 2's guilt in G_5 has the dimension of (expected) material payoff of player 1. And in BD&S model, player i's frustration has the dimension of (expected) material payoff of i. This is *not* a necessary feature of p-utility,and alternatives have been considered. Attanasi, Rimbaud, Villeval (2019) (AR&V) consider "situations where donors need intermediaries to transfer their donations to recipients and where donations can be embezzled before they reach the recipients." They discuss how intermediaries may experience guilt if they do not meet the owner's expectation, although the associated material cost would be incurred by the recipient rather than the donor. And BD&S (in their discussion section) mention how in principle frustration may depend on regret of a previous decision, unexpected perceived unfairness, or negative shocks to self-esteem.

[56] López-Pérez (2008) is an important exception. His model is not PGT-based however.

[57] We do not expect the topic to be easy to address. There are many subtle issues. Is a norm a strategy or a strategy profile? If people like to follow norms, what exactly is the nature of the preference involved? Is the cost of breaking a norm dependent on whether and how many others do so?

Unawareness Almost all game-theoretic analysis assumes that the game form is commonly known between the players. Casual observations of reality suggest that this assumption may be too strong. It seems plausible that before the "Fosbury Flop" and the "V-style" were popularized by, respectively, Dick Fosbury and Jan Boklöv, many high jumpers and ski jumpers were not aware of these techniques, or at least of their lucrative payoff consequences. Or, awareness of the possibility and expected nature of hi-jackings may have been altered by the 9/11 events.

A fairly recent game-theoretic literature develops techniques for modeling unawareness in games. See Heifetz, Meier & Schipper (2006) for a pioneering effort (and Burkhard Schipper provides an "Unawareness bibliography" with further references on his homepage at UC Davis). It is very natural to imagine that belief-dependent motivation interacts with unawareness. For example, one may imagine that negative surprises that reveal previously unforeseen danger could instill fear, so if 9/11 involved unawareness then the occurrence of that event might have consequences for subsequent demand for air travel or supply of airport security.

Exploring unawareness using PGT seems potentially interesting, but we only know one paper that is devoted to the topic: Nielsen & Sebald (2017). We quote the verbal example with which they open their paper (pp. 2-3) as it illustrates nicely how unawareness may interact with a belief-dependent motivation (namely, guilt):[58]

> Assume it is Bob's birthday, he is planning a party and would be very happy, if Ann could come. Unfortunately Bob's birthday coincides with the date of Ann's final exam at university. She can either decide to take the exam the morning after Bob's party or two weeks later at a second date. Ann is certain that Bob would feel let down, if she were to cancel his party without having a very good excuse. Quite intuitively, although Ann would really like to get over her exam as soon as possible, she might anticipate feeling guilty from letting down Bob if she canceled his party to take the exam the following morning. As a consequence, Ann might choose

[58] We notice, however, that equilibrium concepts that do not capture strategic reasoning, such as self-confirming equilibrium, are consistent with some forms of unawareness. For example, players' whose utility depend on their action and a state variable aggregating the actions of other may hold conjectures (confirmed in equilibrium) about such state variable being unaware of the actions of others. See Battigalli, Panebianco, & Pin (2018).

the second date to avoid letting Bob down. In contrast, consider now the following variant of the same example: Ann knows that Bob is unaware of the second date. In this situation Ann might choose to take the exam on the first date and not feel guilty. Since Bob is unaware of the second date and the final exam is a good excuse, he does not expect Ann to come. Ann knows this and, hence, does not feel guilty as Bob is not let down. In fact, if she were certain that Bob would never become aware of the second date, she probably had an emotional incentive to leave him unaware in order not to raise his expectations.

Motivated beliefs The 2016 summer issue of the *Journal of Economic Perspectives* contains an interesting symposium on "Motivated Beliefs," with an introduction by Epley & Gilovich (E&G, who credit George Loewenstein for taking "the leading role in stimulating and organizing the papers") and contributions by Bénabou & Tirole; Golman, Loewenstein, Moene & Zarri; and Gino, Norton & Weber. The idea is this: Beliefs affect people's well-being. This, in turn, affects how they reason, control information, and gather & evaluate evidence. To some extent, it is argued, they may even *choose* their beliefs, although such choice may be unconscious and the ability to do so is hampered by reality-checks and various costs of having faulty beliefs. E&G mention how the topic has "a long history in psychological science" (p. 139). A particularly important reference would seem to be Kunda (1990), who wrote a highly influential paper on how motivation influences reasoning.

It is interesting to reflect on whether and how PGT may be useful in this connection. First, there is overlap on relevant topics. PGT is obviously useful for describing how beliefs affect well-being; such links are embodied in almost every example of belief-dependent utilities that we have exhibited. Second, relatively little work in the literature on motivated beliefs has been math-based, and PGT may provide relevant tools for scholar who want to develop theory. Third, PGT is well equipped to deal with how belief-dependent motivation may impact how people control information, and how they gather evidence. These aspects concern *choices* that presumably can be straightforwardly described in carefully selected game forms. To see this more clearly, note that PGT models the (rational) choice of an agent as a process that takes as *given* his *system of conditional beliefs*, but the *actual* beliefs held on the realized path may well depend on his actions (as well as

actions of others and exogenous shocks).[59] For example, an agent with imperfect recall may store and recall, possibly at a cost, the flow of information he receives, thus manipulating what he is able to remember and his beliefs.[60] For the remaining aspects (modes of reasoning, evaluating evidence, unconscious manipulation of beliefs) it seems less clear to what extent and how PGT provides useful tools. But we are optimists and conjecture that PGT might prove useful for doing that too.

6 Experiments

Theories formulated using PGT can be tested for empirical relevance in lab experiments. We now describe some existing work and reflect on things that could be done. Our focus is on methods of particular relevance to PGT more than on describing results. Also, we focus on experiments that take the PGT-part seriously, rather than just mention some theory is passing as being loosely relevant.[61]

Belief elicitation Models formulated using PGT suggest ways that particular beliefs impact preferences and play, so to conduct lab tests it is often helpful to elicit those beliefs. The very first experiment specifically designed to test a PGT-based prediction was built around that insight. Dufwenberg & Gneezy (2000) (D&G) considered versions of G_4 (recall: player 2 chooses $t \in \{0, ..., M\}$) as well as trust games where 1 could take an outside option or let 2 choose in a subgame structured like G_4 ("lost wallet games"). D&G measured 1's first order-belief (FOB = expectation of t) by asking 1 to *guess* t (with rewards for accuracy). And they measured 2's second-order belief (SOB = the conditional expectation of 1's FOB) by asking 2 to *guess*

[59] Indeed, we touch on examples of this sort, e.g. in Section 4.4 on self-esteem, and also where we discussed the impact of different information structures (Section 2's "third" observation, and the part on "higher-order belief-dependence" of this section).

[60] Compare with Bénabou & Tirole (2002) and their citation from Darwin (1898), where the great scientist describes how he manipulates memory of unpleasant facts to counteract unconscious removal.

[61] For example, hundreds of experimental studies will loosely discuss how reciprocity might be relevant for subjects' decision, and give a reference to D&K in that connection. We do not discuss that literature. By contrast, we cite Dhaene & Bouckaert's (2010) study which set out with the explicit goal of testing D&K's theory and then collected precisely what data they needed for that purpose (including eliciting particular conditional beliefs).

1*'s guess* (again with rewards for accuracy).[62] D&G's test for guilt checks whether for subjects in the position of player 2 there is positive correlation between t and those guess-guesses.

There is a large follow-up literature testing for the empirical relevance of guilt in various games, and which often elicits beliefs. Most commonly, binary trust games like G_5 are explored. See Cartwright (2019) for a survey.[63]

There are also some studies that elicit beliefs in order to study other forms of motivation than guilt. The pioneer to do this for reciprocity theory is Dhaene & Bouckaert (2010),[64] and a few recent studies testing aspects of BD&S' models of frustration and anger also do it.[65]

There are many thorny methodological issues surrounding how to best measure subjects' beliefs.[66] Different PGT-related papers take different approaches and some (e.g. Cartwright) discuss pros & cons. See Schotter & Trevino (2014) for a (broader than just PGT) critical survey of the literature on belief elicitation in laboratory experimental economics.

Belief disclosure C&D point out that the guilt hypothesis just discussed is confounded by a form of "false consensus," if 2's choice (done for whatever reason) shapes her SOB such that she believes others believe she made that

[62] This description is precise as regards G_4. In D&G's lost wallet games, player 2 was actually asked about the *average guess of all the subjects in the role of* 1 *who chose In*. This is crucial to make sure that the right belief is elicited, namely 2's belief conditional on 1 choosing *In*.

[63] See also Guerra & Zizzo (2004), C&D (2006, 2010, 2011), Bacharach, Guerra & Zizzo (2007), Vanberg (2008), Miettinen & Suetens (2008) Reuben, Sapienza & Zingales (2009), Ellingsen, Johannesson, Tjøtta & Torsvik (2010), Bellemare, Sebald & Strobel (2011), Chang, Smith, Dufwenberg & Sanfey (2011), DG&HS, Amdur & Schmick (2013), AB&N, Beck, Kerschbamer, Qiu & Sutter (2013), Bracht & Regner (2013), Kawagoe & Narita (2014), Morrell (2014), Regner & Harth (2014), AG&T, KO&W, Khalmetski (2016), Woods & Servatka (2016), B&F, Balafoutas & Sutter (2017), Bellemare, Sebald & Suetens (2017) (BS&S), Attanasi, Battigalli, Manzoni & Nagel (2019) (ABM&K), AR&V, DK&S, DW&aN) Di Bartolomeo, Dufwenberg, Papa & Passarelli (2019), and Inderst, Khalmetski & Ockenfels (2019).

[64] See also DG&HS, AB&N, and ABM&K.

[65] See AB&G and Dufwenberg, Li & Smith (2018*a*,*b*). Also Persson (2018) tests BD&S theory, although he does so without eliciting beliefs.

[66] For example, should guesses be done before or after choices are made; refer to probabilities of a particular co-player's choices or frequencies of choices among a set of subjects one might be matched with; be incentivized or not, and if so how? These questions often have no obvious answers (for example, a quadratic scoring rule may provide precise incentives to reveal a particular expectation, but may also be harder for a subject to understand).

choice. This would imply that a subject's choice drives his SOB, rather than the other way around (as the guilt story has it). Ellingsen, Johannesson, Tjøtta & Torsvik (2008) (EJT&T) propose a clever alternative design, which avoids that issue but which has another problem. Rather than elicit 2's SOB they elicit 1's FOB, which they then *disclosed* to 2 before she made her choice. This procedure induces 2's SOB without the risk of false consensus. The drawback, however, is a potential loss of control. In EJT&T's design 2 is informed that 1 was not informed that his elicited belief would be handed down to 2. This design feature is important, because if 1 knew then he would have had an incentive to lie (if he believed 2 believed him). The problem is that when 2 learns that some design information is withheld from the players she may wonder if possibly there are other design aspects that are withheld from her. Perhaps that affects her behavior.[67]

No elicitation In many cases it is not necessary to elicit beliefs to meaningfully test PGT-based hypotheses. Sometimes patterns of behavior are idiosyncratic enough to a specific theory that clear conclusions can be drawn by observing choice data alone. For cases in point, consider C&D's (2011) tests regarding "guilt-from-blame" (noting especially their remark at the top of p. 1231); DS&VE's test of whether negative reciprocity plays a role in hold-up problems; tests concerning D&D's theory that manipulate information across end nodes as described under the "third" and "fourth" observation in Section 2; or tests that involve one-player games where the relevant beliefs are pinned down by moves by nature—examples include tests of K&R's theory as pioneered by Ericson & Fuster (2007) (E&F) and by Smith (2019, but written contemporaneously with E&F) and Persson's (2018) test of BD&S.

Communication C&D argued that guilt can help explain why *communication*, and in particular *promises*, can foster trust & cooperation – recall observation (iv) in the guilt part of Section 4.2. They designed an experiment to test that hypothesis, using tests similar to those of D&G described above. Vanberg (2008) argued that C&D's results are confounded by another "commitment-based theory," i.e., that decision makers have a belief-independent preference not to break a promise they made. To test his theory,

[67] This line of criticism has made EJT&T's approach controversial, and yet the technique has come to be frequently relied on. See e.g. AB&N, KO&W, BS&S, ABM&N, DW&aN, and DK&S.

Vanberg came up with an ingenious design, based on a "switching feature." Any subject to whom a pre-play promise were issued was "switched" and replaced by another subject who would play with the person who issued the promise. If there were a switch, the promisor was told but the promisee was not. The key idea is that promisors would suffer expectations-based guilt independently of whether or not a switch occurred, whereas any cost of breaking a promise would apply only if no switch took place. Note that the commitment-based theory is *not* PGT-based. However, discussions of it typically involve comparisons with C&D's belief-based account, so it is important for PGT-scholars to know about Vanberg's work.

Exogeneity & causal inference Vanberg's approach is important also for the following methodological reason: Testing for belief-dependent preferences by comparing subjects who self-report different beliefs, as C&D did, has the drawback of not relying on exogenously created variation. Subjects are not randomly assigned to their beliefs. This weakens the force with which valid causal evidence can be drawn. Similarly, if subjects can choose which message to send, then they are not randomly assigned to their messages. Vanberg overcame this last issue via his switching mechanism, creating exogenous variation is whether or not a subject had sent a promise to the player he eventually interacted with. Vanberg did not attempt to create exogenous variation in subjects SOB though, so his design is not ideal for reconsidering C&D's hypotheses. Ederer & Stremitzer (2017) developed a design that involves exogenous variation in subjects' SOB's, and Di Bartolomeo, Dufwenberg, Papa & Passarelli (2019) developed a design that features exogenous variation in both SOB's and promises. We refer to these studies for more information, while noting that exogenous variation and causal inference has become of high importance in this literature.

Avoidance For certain PGT-related testing purposes it may be useful to employ designs that allow a subject to avoid making another subject aware of a game being played. This would presumably be useful if one were to test ideas that directly involved unawareness, like in C&N's "party example" of section 5, but it can also be useful for testing whether subject care about image as described in section 4.3. For example, consider the design of Dana, Cain & Dawes (2006) (DC&D): Subject i were given a choice whether to "exit" a \$10 dictator game (like G_4, with $M = 10$) and take \$9 instead,

knowing that the exit option would leave the receiver j nothing *and* ensure that j never knew that a dictator game could have been played. DC&D's design provides a test whether i cares for his image. The idea is that by exiting i may enjoy a (rather) high payoff without suffering a bad image.[68]

Alternative design are conceivable that can test similar hypotheses without leaving any player unaware of some strategic possibility, but rather allow a player to avoid revealing a specific choice. For example, the design of Andreoni & Bernheim (2009) examines "an extended version of the dictator game in which (a) nature sometimes intervenes, choosing an unfavorable outcome for the recipient, and (b) the recipient cannot observe whether nature intervened" (p. 1609).[69] The idea is that by choosing directly the "unfavorable outcome" that nature might implement a subject can avoid a bad image.

Other forms of data It may be useful to consider other kinds of data than choices and elicited beliefs to test PGT-based hypotheses. For example, brain imaging data (e.g. fMRI), emotion self-reports ("please rate how strongly you feel emotion X on a scale..."), electrodermal activity, or face-reader data may be useful. Chang, Smith, Dufwenberg & Sanfey (2011) (CSD&S) pioneered the use of fMRI for PGT-related purposes, in a study taking B&D's (2007) theory of simple guilt to the brain scanner. CSD&S' study also involved emotion self-reports, in a way that was mindful of the possibility that pangs of guilt might be counterfactual and yet crucial (compare observation (x) in the guilt part of Section 4.2 above).[70] We do not know of any face-reader study which was conducted with an explicit PGT-connection in mind, but van Leeuwen, Noussair, Offerman, Suetens, van Veelen & van de Ven (2018) (LNOV&V) use the technology to explore anger and BD&S cite LNOV&V's results when motivating their own theory.

[68] Note that an exit choice gives \$-payoff combination $(9, 0)$ to i and j, whereby i reveals a preference for that outcome over each of the \$-payoff combinations of $(10, 0)$, $(9, 1)$, and $(5, 5)$ which he could have obtained by not exiting. This contradicts many models of distributional social preferences, like F&F, B&O, and C&R.

[69] For other related designs, see Dana, Weber & Kuang (2007), Broberg, Ellingsen & Johannesson (2007), and Lazear, Malmendier & Weber (2012).

[70] CSD&S write (p. 569): "To confirm that participants were actually motivated by anticipated guilt, we elicited their counterfactual guilt for each trial following the scanning session. After displaying a recap of each trial, we asked participants how much guilt they would have felt had they returned a different amount of money."

7 Comments on Methodology

Fictitious auxiliary games Can and should we study the psychological phenomena we have described using standard game theory (GT) instead of PGT? Sometimes p-games, most notably perhaps those involving image concerns as described in Section 4.3, can be turned into "strategically equivalent" standard games by endowing the observer with a fictitious action space whereby he reports a belief, or estimate of $\theta_i^{\mathbf{I}}$ and is rewarded with an incentive compatible scoring rule. The receiver's belief—or estimate—is then replaced by his action/report in the sender's utility function. For example, in (12) is replaced by $a_j \geq 0$, j's (pseudo-) utility is

$$u_j\left(\theta_i^{\mathbf{I}}, a_j\right) = -\left(\theta_i^{\mathbf{I}} - a_j\right)^2 \tag{13}$$

and $\mathbb{E}\left[\widetilde{\theta_i^{\mathbf{I}}} | \mathcal{P}_j\left(z\right); \alpha_j\right]$ is replaced by a_j in (12). This works because j maximizes the expected value of (13) by letting a_j equal the conditional expectation of $\widetilde{\theta_i^{\mathbf{I}}}$. As long as i believes in j's rationality, the strategic analysis of the p-game such associated standard game are equivalent.

As in many fields of pure and applied math, transforming a problem into an "equivalent" one may give access to the application of known techniques and results.[71] However, the possibility of such transformations has also engendered the claim that PGT is, after all, not needed: choosing different "weird" assumptions about utility[72] one can go back to good, old, familiar GT, making everybody feel at home. We are very critical of such attitudes. They confuse formalism with reality. The reality is given by the *true* game form (something that can be designed and controlled in the lab) and the true utility (which—in so far as it exists—one can try to elicit under appropriate auxiliary assumptions). If player j is passive in the true game form, coming up with a false representation of reality to claim representability with an old framework can be misleading.[73]

[71] To mention non-obvious ones, results about forward-induction reasoning and rationalizability in a class of infinite dynamic games (Battigalli & Tebaldi 2018) can be applied to p-games with image concerns.

[72] As one colleague and friend of us put it.

[73] Furthermore, nobody has shown that all interesting forms of p-games can be turned into "equivalent" standard games. Considering claims made at seminars attended of presented by us we suspect that this is not for lack of trying. The only article we know of dealing with the topic is that of Kolpin (1992). He limits attention to the class of p-games

For example, the suggestion that standard GT can model everything (possibly after some transformation), even if correct (which we doubt), distracts from important phenomena that cannot be explained when the *true* game is *not* standard, such as the dependence of behavior on the information another player gets when he is inactive. More generally, we propose that adopting indirect techniques referring to known different fields may prevent the conceptually and mathematically correct understanding of the object of analysis, without giving significant benefits over the use of a direct approach. One of the problems is that the theorist often does not realize that he is relying on assumptions that may be wrong/unappealing in some contexts. We illustrate this with a PGT example. GP&S give indirect definitions of different kinds of equilibrium in p-games[74] by first fixing a profile of (initial) belief hierarchies, then looking at equilibria of the resulting standard game and the induced degenerate belief hierarchies whereby beliefs of all orders are correct, and finally defining psychological equilibria as fixed points of the resulting correspondence from the space of profiles of belief hierarchies to itself. This is a clever first step, but it is also convoluted. Indeed, B&D show that a direct definition is both feasible and more transparent. Furthermore, when GP&S consider trembling-hand perfect equilibria (THPE), their indirect approach induces, in our view, a conceptual error: They "forget" that—in the spirit of Selten (1975)—a THPE of the p-game should be limit of equilibria of perturbed p-games whereby players are constrained to "tremble" with strictly positive, although vanishing, minimal probabilities. Using this conceptually correct definition, one can show that a THPE always exists under continuity of utility in beliefs.[75] GP&S instead obtain with their indirect definition a different equilibrium concept and show by example that it may fail to exist (p. 73). What went wrong? Using their indirect approach that considers equilibria of auxiliary standard games, GP&S did not realize that trembles in actions must induce corresponding perturbations in beliefs, hence in the utility of terminal nodes. Once this is taken into account, their counterexample vanishes.

considered by GP&S. In our view, while his exercise is pioneering and useful as an attempt at proof-of-concept, the specific assumptions he engages are too convoluted to be practical.

[74] Of the restricted class they analyze.

[75] Essentially, B&D prove the existence of THPE, to obtain existence of sequential equilibrium, a slightly weaker concept.

Fictitious auxiliary p-games We should note that a convenience-in-modeling argument may cut in the other directions as well: Models with an intrinsic concern for belief-dependent reputation can sometimes offer a compact reduced-form approach to modeling repeated interaction in settings where players are not assumed to have any belief-dependent motivation. That is, a p-game can in such a case be useful tool for analyzing a standard game. For examples that take such an approach, see Morris (2001) and Ottaviani & Sorensen (2006).

Solution concepts PGT-analysis involves two key steps: (i) modeling belief-dependent utility, and (ii) applying a solution concept. Step (i) is unique to p-games, step (ii) is relevant also for traditional games. Our main goal has been to emphasize what is unique to p-games, so we have focused mostly on step (i). However, since we feel strongly about step (ii), let us explain our view:

Sadly, economists have been socialized to uncritically take for granted that ad hoc notions of equilibrium (whereby players are assumed to have correct beliefs) meaningfully describe strategic interaction. In one-shot play settings, if players reason about each other's rationality and beliefs, inferences should concern steps of deletion of non-best-replies (possibly all the way to "rationalizability"). If learning is allowed by looking at recurrent interaction, the appropriate solution is (some version of) self-confirming equilibrium (SCE), in which beliefs may be incorrect, although consistent with evidence. In neither case is the most commonly applied solution—sequential equilibrium (SE)—generally implied.[76] Only in rare cases is assuming SE justified, in particular when SE is equivalent to rationalizability or SCE.[77]

Since a proper discussion would call for its own article, we have not gone there. Our approach has mainly been consistent with our favored view as we focused on steps of deletion of non-best-replies. But since previous scholarship (including ours) often referred to more traditional notions of equilibrium,

[76] B&D extend Kreps & Wilson's (1982) classic notion of sequential equilibrium to p-games. See BC&D for relevant p-games definitions of all solution concepts mentioned above. See Battigalli, Corrao & Sanna (2019) and Jagau & Perea (2018) for epistemic foundations of (versions of) rationalizability.

[77] A more special circumstance may apply in D&D's model, presented in Section 2, given the interpretation that player 2 is player 1's "imagined" audience (as hinted at). If 1 is, so-to-say, "his own audience," we have a one-player game, so forming equilibrium expectations should be easy" (D&D, p. 262).

we made a few related references when recalling such work.

We hope future work will take the appropriateness and relevance of solution concepts more seriously than has been done in the past.

Transformations of the game form A small literature studies transformations of the game form (with monetary payoffs)[78] that do not change the reduced normal form (see in particular Thompson 1952, Dalkey 1953, and Elms & Reny 1994). Considering only those that preserve perfect recall, they are: interchanging essentially simultaneous moves (INT),[79] coalescing sequential moves (COA; inverse: sequential agent splitting), and addition of a superfluous move (ADD). Any solution concept that just requires knowledge of the normal form, such as Nash equilibrium, or iterated admissibility, is obviously invariant to such transformations (and their inverses). Yet, such invariance cannot be a dogma. Note that the most used solution concept, sequential equilibrium, is only invariant to INT. Versions of rationalizability for extensive-form games are invariant to INT and COA, but not ADD.[80] Our general view is that it is interesting to study the invariance properties of independently motivated solution concepts, without viewing violations of some form of invariance as a flaw. We have just one *caveat*: If simultaneity of moves is represented indirectly, by letting players move in an arbitrary sequence and defining information sets so that the choices of early movers are not observed, then invariance w.r.t. INT is a must.[81]

How does this relate to the analysis of p-games? We have a two-tiered answer:

(i) In general p-games, players' utility depends on the *temporal* sequence of beliefs. This means that an accurate representation must explicitly account for time and distinguish between periods and stages within periods, as–say– in models of bargaining protocols. Transformations involving moves across periods should be expected to have an impact on behavior. For example, as shown by Gneezy & Imas (2014), angry agents "cool off" as time goes by and

[78] The literature refers to "extensive-form games," which include a specification of players' payoffs at terminal nodes. We interpret such payoffs as monetary one, hence we have what we call "game forms".

[79] Note that the formalism initially adopted to describe games indirectly represented simultaneous moves in an arbitrary sequential order without flows of information to late movers.

[80] See the discussion and characterization by Battigalli, Leonetti & Maccheroni (2019).

[81] See the characterization of INT-invariance by Bonanno (1992).

are therefore less willing to harm others (see the discussion of fast *vs* slow play in the working paper version of BD&S). Consider a game form with time in which player i in period t (when he may be frustrated) decides to stop (S) or continue (C). If C is chosen, the play moves to period $t+1$ when i can harm j (H) or not (N). Since i in period $t+1$ may have cooled off, a COA transformation making i choose either S, or $C\&H$, or $C\&N$ in period t, makes more likely that i harms j. Thus, it is important to use rich and accurate representations of the rules of the game (see BC&D for a first step in this direction). The rules of games played in the field and in the lab, by necessity, imply that information accrues to both active and inactive players. Since information affects beliefs, we should formally represent the information of inactive players. Also, the rules of the game only specify information *flows*, not how much information is retained by players. To keep a sharp separation between game form and players' personal features, the former should specify information flows as, for example, in Myerson (1986), or in the literature on repeated games with imperfect monitoring (e.g., Mailath & Samuelson 2006). The use of traditional information partitions is acceptable under the presumption that players have perfect recall (and this is common knowledge).

(ii) Compared to standard game theory, invariance with respect to a particular transformation depends also on the form of psychological utility functions, not only on the solution concept. For example, if utility depends only on the initial, or the terminal beliefs of others, sequential equilibrium is invariant w.r.t. INT (which, however, may not be applicable if simultaneity is represented directly as we do), whereas extensive-form rationalizability is invariant to both INT and COA. Yet, utility may depend on beliefs in ways that prevent any form of invariance.

Private sensitivities and incomplete information In preceding sections θ_i often denoted player i's "sensitivity" with respect to some psychological concerns (e.g. guilt or reciprocity). It makes sense to assume that player j is not informed of θ_i, and has to form beliefs about it in order to predict i's behavior. In this case the analysts must consider beliefs in p-games with incomplete information. We already discussed this topic, in Section 4.3, where, however, we focused on image concerns, i.e. how i might care about j's beliefs about θ_i. However, even absent image concerns, incomplete information is very natural and important. If θ_i measures i's sensitivity to feel, e.g., guilt, reciprocity, or anger, then j's beliefs about θ_i may matter not

because i cares about those beliefs but because they will impact j's choices.[82]

Outside of the image-concerns literature, most applied work using PGT either assumes complete information, or (legitimately) ignores the issue by focusing on the shape of the best reply correspondences, without using solution concepts. The analysis of psychological games with incomplete information is developed in AB&M, Bjorndahl, Halpern & Pass (2019) and BC&D, which is more general and systematic. Incomplete information is the lack of common knowledge of some features of the game form (feasible sets, information structure, material payoff functions) or of players' preferences. As explained in BC&D, if we rely on solution concepts with an epistemic foundation (versions of rationalizability) or a learning foundation (versions of self-confirming equilibrium), incompleteness of information can be addressed directly by means of relatively straightforward extensions of such solution concepts.[83] The situation is different if the analyst maintains the traditional equilibrium assumption that players have correct conjectures about the decision rules associating co-players' private information with their behavior. In this case, following the seminal contribution by Harsanyi (1967-68), the traditional approach proceeds in two steps. It first posits an *implicit* representation of players' possible interactive beliefs about the unknown parameter vector θ by means of "types" encoding both private knowledge about θ and hierarchical exogenous beliefs; this yields a "Bayesian game". Then one proceeds to analyze the Bayesian perfect (or sequential) equilibria of such game, that is, profiles of decision rules whereby every type of every player carries out a sequential best reply to the co-players' decision rules given the beliefs of this type. The key to connect Harsanyi's approach to PGT is that, in a Bayesian equilibrium, each type of each player is associated to a hierarchy of (exogenous *and* endogenous) beliefs, which—in a p-game—enters the utility functions. This allows to obtain all the known "traditional" equilibrium concepts (those that postulate correct beliefs, as in GP&S and B&D) as special cases of Bayesian perfect equilibrium. Yet, Harsanyi's approach gives an additional degree of flexibility seldom noticed by non-specialists: a Bayesian game may feature multiple types of the same player with the *same exogenous* hierarchy of beliefs, even when there is complete information! This allows for equilibria where types with the same exogenous belief hierarchy have dif-

[82] For analysis of games where players are uncertain about each others' such sensitivities, see S&W on reciprocity; AB&M, ABM&N, ABM&N, and BCh&D (footnote 5) on guilt; and Aina, Battigalli & Gamba (2018) (AB&G) on anger.

[83] See Sections 7.1-2 of BC&D and the references therein.

ferent plans and different beliefs about co-players. Since observed actions signal types, in equilibrium players may update their higher-order beliefs about the beliefs (and plans) of co-players, hence their intentions, which is instead prevented by the equilibrium concepts of GP&S and B&D.[84]

Bounded rationality All the work on PGT that we know of assumes that players are rational. Clearly, however, the idea of bounded rationality makes as much sense in the context of p-games as in standard games. We propose that exploring related topics might be interesting, although we mainly leave exploring it for future efforts. We see at least one interesting potential interaction between bounded rationality and PGT: While the latter models emotions as part of players' utility, it is known that some emotions such as anxiety (Rauh & Seccia, 2006) and anger (Gneezy & Imas 2014) can hamper rational cognition, and this is factored in by early movers who can trigger such emotions. PGT in its current form is not equipped to model such effects.

8 Concluding remarks

Decisions are driven by a plethora of desires. Yet economists' approaches traditionally took a narrow view, focusing mainly on concern for own income (or consumption). When richer models were proposed, it was often taken as an advantage if the deviations from the tradition were limited. For example, much of the literature on "social preferences" considers it a success if data sets can be explained using utilities defined on distributions of material payoffs according to simple formulas.[85]

Being spare is not necessarily a virtue. If human psychology is rich and multi-faceted, one cannot know the effect of the involved sentiments unless one dives in and explores how and why that plays out in economic contexts. Many interesting desires that shape behavior in important ways take the form of belief-dependent motivation. This includes reciprocity, emotions, image concerns, and self-esteem. We have argued that the mathematical framework of psychological game theory (PGT) is useful and needed for modeling such sentiments, and we have tried to shown why & how. Working with PGT is

[84] See Section 7.3 of BC&D.

[85] See e.g. F&S, B&O, C&R for models, and Cooper & Kagel (2009) for a survey in that spirit.

exciting and we derive utility from our *hope* (=item #12 in Elster's list) to inspire others to follow suit.

References

[1] Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for Truth-Telling". *Econometrica* 87: 1115-53.

[2] Aina, Chiara, Pierpaolo Battigalli, and Astrid Gamba. 2018. "Frustration and Anger in the Ultimatum Game: An Experiment". Bocconi University IGIER Working Paper 621.

[3] Akerlof, George. 1982. "Labour Contracts as a Partial Gift Exchange". *Quarterly Journal of Economics* 97: 543-69.

[4] Aldashev, Gani, Georg Kirchsteiger, and Alexander Sebald. 2017. "Assignment Procedure Biases in Randomized Policy Experiments". *The Economic Journal* 127: 873–895.

[5] Amdur, David, and Ethan Schmick. 2013. "Does the Direct-Response Method Induce Guilt Aversion in a Trust Game?". *Economics Bulletin* 33(1): 687–693.

[6] Andreoni, James, and B. Douglas Bernheim. 2009. "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects". *Econometrica* 77: 1607-1636.

[7] Andrighetto, Giulia, Daniela Grieco, and Luca Tummolini. 2015. "Perceived Legitimacy of Normative Expectations Motivates Compliance with Social Norms when Nobody is Watching." *Frontiers in Psychology* 6.

[8] Attanasi, Giuseppe, Pierpaolo Battigalli, and Elena Manzoni. 2016. "Incomplete Information Models of Guilt Aversion in the Trust Game". *Management Science* 62: 648-667.

[9] Attanasi, Giuseppe, Pierpaolo Battigalli, Elena Manzoni, and Rosemarie Nagel. 2019 (forthcoming). "Belief-Dependent Preferences and Reputation: Experimental Analysis of a Repeated Trust Game". *Journal of Economic Behavior and Organization* .

[10] Attanasi Giuseppe, Pierpaolo Battigalli, and Rosemarie Nagel. 2013. "Disclosure of Belief-Dependent Preferences in the Trust Game". Bocconi University IGIER Working Paper 506.

[11] Attanasi, Giuseppe, Claire Rimbaud, and Marie-Claire Villeval. 2019 (forthcoming). "Embezzlement and Guilt Aversion". *Journal of Economic Behavior and Organization.*

[12] Averill, James R. 1982. *Anger and Aggression: An Essay on Emotion.* New York: Springer.

[13] Azar, Ofer H. 2019 (forthcoming). "The Influence of Psychological Game Theory". *Journal of Economic Behavior and Organization.*

[14] Bacharach, Michael, Gerardo Guerra, and Daniel J. Zizzo. 2007. "The Self-Fulfilling property of Trust: an Experimental Study". *Theory and Decision* 63(4): 349–388.

[15] Balafoutas, Loukas. 2011. "Public Beliefs and Corruption in a Repeated Psychological Game". *Journal of Economic Behavior and Organization* 78: 51-59.

[16] Balafoutas, Loukas, and Helena Fornwagner. 2017. "The Limits of Guilt". *Journal of the Economic Science Association* 3(2): 137–148.

[17] Balafoutas, Loukas, and Matthias Sutter. 2017. "On the Nature of Guilt Aversion: Insights from a New Methodology in the Dictator Game". *Journal of Behavioral and Experimental Finance* 13: 9–15 .

[18] Battigalli Pierpaolo, Gary Charness, and Martin Dufwenberg. 2013. "Deception: The Role of Guilt". *Journal of Economic Behavior and Organization* 93: 227-232.

[19] Battigalli Pierpaolo, Roberto Corrao, and Martin Dufwenberg. 2019 (forthcoming). "Incorporating Belief-Dependent Motivation in Games". *Journal of Economic Behavior and Organization.*

[20] Battigalli Pierpaolo, Roberto Corrao, and Federico Sanna. 2019. "Epistemic Game Theory without Types Structures: An Application to Psychological Games". Bocconi University IGIER Working Paper 641.

[21] Battigalli, Pierpaolo, and Nicodemo De Vito. 2018. "Beliefs, Plans, and Perceived Intentions in Dynamic Games". Bocconi University IGIER Working Paper 629.

[22] Battigalli, Pierpaolo, and Martin Dufwenberg. 2007. "Guilt in Games". *American Economic Review* 97(2): 170-176.

[23] Battigalli, Pierpaolo, and Martin Dufwenberg. 2009. "Dynamic Psychological Games". *Journal of Economic Theory* 144: 1-35.

[24] Battigalli, Pierpaolo, Martin Dufwenberg, and Alec Smith. 2019 (forthcoming). "Frustration, Aggression and Anger in Leader-Follower Games." *Games and Economic Behavior* 117, 15-39.

[25] Battigalli, Pierpaolo, Paolo Leonetti, and Fabio Maccheroni. 2019. "Behavioral Equivalence of Extensive Game Structures". Unpublished.

[26] Battigalli, Pierpaolo, Fabrizio Panebianco, and Paolo Pin. 2018. "Learning and Self-Confirming Equilibrium in Network Games". Bocconi University IGIER Working Paper 637.

[27] Battigalli, Pierpaolo, and Marciano Siniscalchi. 1999. "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games". *Journal of Economic Theory* 88: 188-230.

[28] Battigalli, Pierpaolo, and Pietro Tebaldi. 2018 (forthcoming). "Interactive Epistemology in Simple Dynamic Games with a Continuum of Strategies". *Economic Theory.*

[29] Baumeister, Roy F., Arlene M. Stillwell, and Todd F. Heatherton. 1994. "Guilt: An Interpersonal Approach". *Psychological Bulletin* 115(2): 243-267.

[30] Beck Adrian, Rudolf Kerschbamer, Jianying Qiu, and Matthias Sutter. 2013. "Shaping Beliefs in Experimental Markets for Expert Services: Guilt Aversion and the Impact of Promises and Money-Burning Options". *Games and Economic Behavior.* 81(September): 145–164.

[31] Bell, David. 1982. "Regret in Decision Making under Uncertainty". *Operations Research* 30: 961-981.

[32] Bell, David. 1985. "Disappointment in Decision Making under Uncertainty". *Operations Research* 33: 1-27.

[33] Bellemare, Charles, Alexander Sebald, and Martin Strobel. 2011. "Measuring the Willingness to Pay to Avoid Guilt: Estimation using Equilibrium and Stated Belief Models". *Journal of Applied Economics*. 26 (3): 437–453.

[34] Bellemare, Charles, Alexander Sebald, and Sigrid Suetens. 2017. "A Note on Testing Guilt Aversion". *Games and Economic Behavior* 102: 233-239.

[35] Bénabou, Roland, and Jean Tirole. 2002. "Self-Confidence and Personal Motivation". *Quarterly Journal of Economics*, 117: 871-915.

[36] Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior". *American Economic Review* 96: 1652-78.

[37] Bénabou, Roland, and Jean Tirole. 2016. "Mindful Economics: The Production, Consumption, and Value of Beliefs". *Journal of Economic Perspectives* 30: 141-64.

[38] Berglas, Steven, and Edwin E. Jones. 1978. "Drug Choice as a Self-Handicapping Strategy in Response to Noncontingent Success". *Journal of Personality and Social Psychology* 36: 405-417.

[39] Berkowitz, Leonard. 1978. "Whatever Happened to the Frustration-Aggression Hypothesis?". *American Behavioral Scientist* 21: 691-708.

[40] Berkowitz, Leonard. 1989. "Frustration-Aggression Hypothesis: Examination and Reformulation". *Psychological Bulletin* 106: 59-73.

[41] Bernheim, Douglas. 1994. "A Theory of Conformity". *Journal of Political Economy* 102: 841-877.

[42] Bicchieri, Cristina. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge, MA: Cambridge University Press.

[43] Bierbrauer, Felix, and Nick Netzer. 2016. "Mechanism Design and Intentions". *Journal of Economic Theory* 163: 557–603.

[44] Bierbrauer, Felix, Axel Ockenfels, Andreas Pollak, and Désirée Rückert. 2017. "Robust Mechanism Design and Social Preferences". *Journal of Public Economics* 149: 59-80.

[45] Bjorndahl, Adam, Joseph Halpern, and Rafael Pass. 2019. "Bayesian Games with Intentions." Unpublished.

[46] Blume, Andreas, Ernest K. Lai, and Wooyoung Lim. 2019 (forthcoming). "Eliciting Private Information with Noise: The Case of Randomized Response". *Games and Economic Behavior*.

[47] Bolton, Gary, and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition". *American Economic Review* 90: 166-193.

[48] Bonanno, Giacomo. 1992. "Set-Theoretic Equivalence of Extensive-Form Games". *International Journal of Game Theory* 20: 429-447.

[49] Bracht, Jurgen, and Tobias Regner. 2013. "Moral Emotions and Partnership". *Journal of Economic Psychology*. 39: 313–326.

[50] Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson. 2007. "Is Generosity Involuntary?." *Economics Letters* 94(1): 32–7.

[51] Caplin, Andrew, and John Leahy. 2001. "Psychological Expected Utility Theory and Anticipatory Feelings". *Quarterly Journal of Economics* 116: 55-79.

[52] Caplin, Andrew, and John Leahy. 2004. "The Supply of Information by a Concerned Expert". *Economic Journal* 114: 487-505.

[53] Card, David, and Gordon Dahl. 2011. "Family Violence and Football: The Effect of Unexpected Emotional Cues on Violent Behavior". *Quarterly Journal of Economics* 126: 103-143.

[54] Cardella, Eric. 2016. "Exploiting the Guilt Aversion of Others: Do Agents Do It and Is It Effective?". *Theory and Decision* 80: 523-560.

[55] Caria, Stefano, and Marcel Fafchamps. 2019 (forthcoming). "Expectations, Network Centrality, and Public Good Contributions: Experimental Evidence from India". *Journal of Economic Behavior and Organization*.

[56] Cartwright, Edward. 2019 (forthcoming). "A Survey of Belief-based Guilt Aversion in Trust and Dictator Games". *Journal of Economic Behavior and Organization*.

[57] Çelen, Bogaçhan, Andrew Schotter, and Mariana Blanc. 2017. "On Blame and Reciprocity: Theory and Experiments". *Journal of Economic Theory* 169: 62-92.

[58] Cerreia-Vioglio, Simone, David Dillenberger, and Pietro Ortoleva. 2018. "An Explicit Representation for Disappointment Aversion and Other Betweenness Preference". Bocconi University IGIER w.p. 631.

[59] Chang, Luke, Alec Smith, Martin Dufwenberg, and Alan Sanfey. 2011. "Triangulating the Neural, Psychological, and Economic Bases of Guilt Aversion". *Neuron* 70(3): 560-72.

[60] Charness, Gary, and Martin Dufwenberg. 2006. "Promises and Partnership". *Econometrica* 74: 1579-1601.

[61] Charness, Gary, and Martin Dufwenberg. 2010. "Bare promises: an experiment". *Economics Letter* 107(2): 281–283.

[62] Charness, Gary, and Martin Dufwenberg. 2011. "Participation". *American Economic Review* 101: 1213-39.

[63] Charness, Gary, and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests". *Quarterly Journal of Economics* 117: 817-869.

[64] Conconi, Paola, David R. DeRemer, Georg Kirchsteiger, Lorenzo Trimarchi, and Maurizio Zanardi. 2017. "Suspiciously Timed Trade Disputes". *Journal of International Economics* 105: 57-75.

[65] Cooper, David J., and John H. Kagel. 2016. "Other Regarding Preferences: A Survey of Experimental Results". In *The Handbook of Experimental Economics*. Vol. 2. Princeton: Princeton University Press.

[66] Cox, James, Daniel Friedman, and Vjollca Sadiraj. 2008. "Revealed Altruism". *Econometrica* 76: 31-69.

[67] d'Adda, Giovanna, Martin Dufwenberg, Francesco Passarelli, and Guido Tabellini. 2019. "Partial Norms". Bocconi Univeristy IGIER Working Paper 643.

[68] Dalkey, Norman. 1953. "Equivalence of Information Patterns and Essentially Determinate Games". *Contributions to the Theory of Games II*, ed. by H.W. Kuhn and A.W. Tucker. Princeton. Princeton University Press, 217-243.

[69] Dana, Jason, Daylian Cain, and Robyn M. Dawes. 2006. "What You Don't Know Won't Hurt Me: Costly (but Quiet) Exit in Dictator Games." *Organizational Behavior and Human Decision Processes* 100 (2):193–201.

[70] Dana, Jason, Roberto Weber, and Jason Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness." *Economic Theory* 33(1) :67–80.

[71] Danilov, Anastasia, Kyril Khalmetski, and dirk Sliwka. 2019. "Norms and Guilt". CESifo Working Paper Series 6999. CESifo Group Munich.

[72] Dhami, Sanjit, Mengxing Wei, and Ali al-Nowaihi. 2019 (forthcoming). "Public Goods Games and Psychological Utility: Theory and Evidence". *Journal of Economic Behavior and Organization.*

[73] Dhaene, Geert, and Jan Bouckaert. 2010. "Sequential Reciprocity in Two-Player, Two-Stage Games: An Experimental Analysis". *Games and Economic Behavior* 70: 289-303.

[74] Di Bartolomeo, Giovanni, Martin Dufwenberg, Stefano Papa, and Francesco Passarelli. 2019. "Promises, Expectations and Causation". *Games and Economic Behavior* 113: 137-46.

[75] Dollard, John, Leonard W. Doob, Neal E. Miller, O. H., Mowrer, and Robert R. Sears. 1939. *Frustration and Aggression.* New Haven: Yale University Press.

[76] Dufwenberg, Martin. 2002. "Marital Investment, Time Consistency and Emotions". *Journal of Economic Behavior and Organization* 48: 57-69.

[77] Dufwenberg, Martin. 2008. "Psychological Games". In *The New Palgrave Dictionary of Economics* edited by S.N. Durlauf and L.E. Blume. Volume 6: 714-18. Palgrave Macmillan.

[78] Dufwenberg, Martin, and Martin A. Dufwenberg. 2018. "Lies in Disguise - A Theoretical Analysis of Cheating". *Journal of Economic Theory* 175: 248-264.

[79] Dufwenberg, Martin, Simon Gächter, and Heike Hennig-Schmidt. 2011."The Framing of Games and the Psychology of Play". *Games and Economic Behavior* 73: 459-478.

[80] Dufwenberg, Martin, Katja Görlitz and Christina Gravert. 2019. "Peer Evaluation Tournaments". Mimeo.

[81] Dufwenberg, Martin and Uri Gneezy. 2000. "Measuring Beliefs in an Experimental Lost Wallet Game." *Games and Economic Behavior* 30: 163-182.

[82] Dufwenberg, Martin, and Georg Kirchsteiger. 2000. "Reciprocity and Wage Undercutting". *European Economic Review* 44: 1069-1078.

[83] Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior* 47: 268-298.

[84] Dufwenberg, Martin and Georg Kirchsteiger. 2019 (forthcoming). "Modelling Kindness". *Journal of Economic Behavior and Organization.*

[85] Dufwenberg, M., Flora Li, and Alec Smith. 2018. "Promises and Punishment". Unpublished.

[86] Dufwenberg, Martin, Flora Li, and Alec Smith. 2018. "Threats". Unpublished.

[87] Dufwenberg, Martin, and Senran Lin, "Regret Games". Unpublished.

[88] Dufwenberg, Martin, and Michael Lundholm. 2001. "Social Norms and Moral Hazard". *Economic Journal* 111: 5.

[89] Dufwenberg, Martin, and Katarina Nordblom. 2018. "Tax Evasion with a Conscience". Unpublished.

[90] Dufwenberg, Martin , and Amrish Patel. 2017. "Reciprocity Networks and the Participation Problem". *Games and Economic Behavior* 101: 260-272

[91] Dufwenberg, Martin, and David Rietzke. 2016. "Banking on reciprocity: deposit insurance and insolvency". Mimeo.

[92] Dufwenberg, Martin, Alec Smith, and Matt Van Essen. 2013. "Hold-up: With a Vengeance". *Economic Inquiry* 51: 896-908.

[93] Ederer, Florian, and Alexander Stremitzer. 2017. "Promises and Expectations". *Games and Economic Behavior* 106: 161-178.

[94] David Eil and Justin M. Rao (2011) Asymmetric Processing of Objective Information about Yourself American Economic Journal: Microeconomics 3 (May 2011): 114–138.

[95] Ellingsen, Tore, Magnus Johannesson, Sigve Tjøtta, Gaute Torsvik. 2010. "Testing Guilt Aversion". *Games and Economic Behavior* 68: 95-107.

[96] Elms, Susan, and Phil Reny. 1994. "On the Strategic Equivalence of Extensive Form Games". *Journal of Economic Theory* 62: 1-23.

[97] Epley, Nicholas, and Thomas Gilovich. 2016. "The Mechanics of Motivated Reasoning". *Journal of Economic Perspectives* 30: 133-40.

[98] Ellingsen, Tore, and Magnus Johannesson. 2008. "Pride and Prejudice: The Human Side of Incentive Theory". *American Economic Review* 98: 990-1008.

[99] Elster, Jon. 1989. "Social Norms and Economic Theory". *The Journal of Economic Perspectives* 3(4): 99-117.

[100] Elster, Jon. 1998. "Emotions and Economic Theory". *Journal of Economic Literature* 36: 47-74.

[101] Ely, Jeffrey, Alexander Frankel, and Emir Kamenica. 2015. "Suspense and Surprise". *Journal of Political Economy* 123, 215-260.

[102] Engelbrecht-Wiggans, Richard. 1989. "The Effect of Regret on Optimal Bidding in Auctions". *Management Science* 35(6): 685-92.

[103] Engelbrecht-Wiggans, Richard, and Elena Katok. 2008. "Regret and Feedback Information in First-Price Sealed-Bid Auctions". *Management Science* 54(4): 808-819.

[104] Ericson, Keith Marzilli, and Andreas Fuster. 2011. "Expectations as Endowments: Evidence on Reference-Dependent Preferences from Exchange and Valuation Experiments". *Quarterly Journal of Economics* 126: 1879-1907.

[105] Falk, Armin, and Urs Fischbacher. 2006. "A Theory of Reciprocity". *Games and Economic Behavior* 54: 293-315.

[106] Fehr, Ernst, and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity". *Journal of Economic Perspectives* 14: 159–181.

[107] Fehr, Ernst, and Ivo Schurtenberger. 2018. "Normative Foundations of Human Cooperation". *Nature* 2: 458-468.

[108] Fehr, Ernst, and Klaus Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation". *Quarterly Journal of Economics* 114: 817-868.

[109] Filiz-Ozbay, Emel, and Erkut Ozbay. 2007. "Auctions with Anticipated Regret: Theory and Experiment". *American Economic Review* 97: 1407-1418.

[110] Fischbacher, Urs, and Franziska Föllmi-Heusi. 2013. "Lies in Disguise - An Experimental Study on Cheating". *Journal of the European Economic Association* 11: 525-547.

[111] Geanakoplos, John. 1996. "The Hangman Paradox and the Newcomb's Paradox as Psychological Games". Cowles Foundation Discussion Paper No. 1128.

[112] Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological Games and Sequential Rationality". *Games and Economic Behavior* 1: 60-80.

[113] Gilboa, Itzhak, and David Schmeidler. 1988. "Information Dependent Games: Can Common Sense be Common Knowledge?". *Economics Letters* 27: 215–221.

[114] Gilboa, Itzhak, and David Schmeidler. 1989. "Maximin Expected Utility with a Non-Unique Prior." *Journal of Mathematical Economics* 18, 141-53.

[115] Gill, David, and Victoria Prowse. 2012. "A Structural Analysis of Disappointment Aversion in a Real Effort Competition". *American Economic Review* 102: 469–503.

[116] Gino, Francesca, Michael I. Norton, and Roberto A. Weber. 2016. "Motivated Bayesians: Feeling Moral While Acting Egoistically". *Journal of Economic Perspectives* 30: 189-212.

[117] Glazer Amihai, and Kai A. Konrad. 1996. "A Signaling Explanation for Charity". *American Economic Review* 86: 1019-1028.

[118] Gneezy, Uri, and Alex Imas. 2014. "Materazzi Effect and the Strategic Use of Anger in Competitive Interactions". *Proceedings of the National Academy of Sciences* 111: 1334-1337.

[119] Gneezy, Uri, Agnel Kajackaite, and Joel Sobel. 2018. "Lying Aversion and the Size of the Lie". *American Economic Review* 108: 419-453.

[120] Golman, Russell, George Loewenstein, Karl Ove Moene and Luca Zarri. 2016. "The Preference for Belief Consonance". *Journal of Economic Perspectives* 30: 165-88.

[121] Goranson, Richard, and Leonard Berkowitz. 1966. "Reciprocity and Responsibility Reactions to Prior Help". *Journal of Personality and Social Psychology* 3: 227-232.

[122] Grossman, Zachary and Joel J. van der Weele. 2017. "Self-Image and Willful Ignorance in Social Decisions". *Journal of the European Economic Association* 15(1): 173–217.

[123] Gradwohl, Ronen, and Rann Smorodinsky. 2017. "Perception Games and Privacy," *Games and Economic Behavior* 104, 293-308.

[124] Guerra, Gerardo, and Daniel J. Zizzo. 2004. "Trust Responsiveness and Beliefs". *Journal of Economic Behavior and Organization* 55(1): 25–30.

[125] Gul, Faruk. 1991. "A Theory of Disappointment Aversion". *Econometrica* 59: 667-686.

[126] Gul, Faruk, and Wolfgang Pesendorfer. 2016. "Interdependent Preference Models As a Theory of Intentions". *Journal of Economic Theory* 165: 179-208.

[127] Hahn, Volker. 2009. "Reciprocity and Voting". *Games and Economic Behavior* 67: 467-480.

[128] Harsanyi, John. 1967-68. "Games of Incomplete Information Played by Bayesian Players. Parts I, II, III". *Management Science* 14: 159-182, 320-334, 486-502.

[129] Heifetz, Aviad, Martin Meier, and Burckard Schipper. 2006. "Interactive Unawareness". *Journal of Economic Theory* 130: 78-94.

[130] Inderst, Roman, Kyril Khalmetski, and Axel Ockenfels. 2019 (forthcoming). "Sharing Guilt: How Better Access to Information May Backfire". *Management Science*

[131] Isoni, Andrea, and Robert Sugden. 2019 (forthcoming). "Reciprocity and the Paradox of Trust in Psychological Game Theory". *Journal of Economic Behavior and Organization.*

[132] Jagau, Stephen, and Andrés Perea. 2018. "Common Belief in Rationality in Psychological Games". Epicenter Working Paper 10.

[133] Jang, Dooseok, Amrish Patel, and Martin Dufwenberg. 2018. "Agreements with Reciprocity: Co-Financing and MOUs". *Games and Economic Behavior* 111: 85-99.

[134] Jiang, Lianjie, and Jiabin Wu. 2019. "Belief-Updating Rule and Sequential Reciprocity". *Games and Economic Behavior* 113: 770-780.

[135] Kahneman, Daniel, and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Risk". *Econometrica* 47: 263-291.

[136] Karni, Edi, and David Schmeidler. 2016. "An Expected Utility Theory for State-Dependent Preferences." *Theory and Decision* 81: 467-478.

[137] Kartik, Navin. 2019. "Strategic Communication with Lying Costs". *Review of Economic Studies* 76: 1359-1395.

[138] Khalmetski, Kyril. 2016. "Testing Guilt Aversion with an Exogenous Shift in Beliefs". *Games and Economic Behavior* 97: 110–119 .

[139] Kawagoe, Toshiji, and Yosuke Narita. 2014. "Guilt Aversion Revisited: an Experimental Test of a New Model". *Journal of Economic Behavior and Organization* 102: 1–9 .

[140] Khalmetski, Kiryl. 2019 (forthcoming). "The Hidden Value of Lying: Evasion of Guilt in Expert Advice". *Journal of Economic Behavior and Organization.*

[141] Khalmetski, Kiryl, Axel Ockenfels, and Peter Werner. 2015. "Surprising Gifts: Theory and Laboratory Evidence". *Journal of Economic Theory* 159: 163-208.

[142] Khalmetski, Kiryl and Dirk Sliwka. 2019 (forthcoming). "Disguising Lies - Image Concerns and Partial Lying in Cheating Games". *American Economic Journal: Microeconomics.*

[143] Kolpin Van. 1992. "Equilibrium Refinements in Psychological Games". *Games and Economic Behavior* 4: 218–231.

[144] Kőszegi, Botond. 2006. "Ego Utility, Overconfidence, and Task Choice". *Journal of the European Economic Association.* 4(4): 673–707.

[145] Kőszegi, Botond. 2010. "Utility from Anticipation and Personal Equilibrium". *Economic Theory* 44: 415-444.

[146] Kőszegi, Botond, George Loewenstein, and Takeshi Murooka. 2019. "Fragile Self-Esteem." Unpublished.

[147] Kőszegi, Botond, and Matthew Rabin. 2006. "A Model of Reference-Dependent Preferences". *Quarterly Journal of Economics* 121: 1133-1166.

[148] Kőszegi, Botond, and Matthew Rabin. 2007. "Reference-Dependent Risk Attitudes". *American Economic Review* 97: 1047-1073.

[149] Kőszegi, Botond, and Matthew Rabin. 2009. "Reference-Dependent Consumption Plans". *American Economic Review* 99: 909-936.

[150] Kozlovskaya, Maria, and Antonio Nicolò. 2019 (forthcoming). "Public Good Provision Mechanisms and Reciprocity". *Journal of Economic Behavior and Organization*.

[151] Kreps, David, and Evan Porteus. 1978. "Temporal Resolution of Uncertainty and Dynamic Choice Theory." *Econometrica* 46: 185-200.

[152] Kreps, David, and Robert Wilson. 1982. "Sequential Equilibria". *Econometrica* 50: 863-894.

[153] Kunda, Ziva. 1990. "The Case for Motivated Reasoning". *Psychological Bulletin* 108: 480-498.

[154] Lazear, Edward, Ulrike Malmendier, and Roberto Weber. 2012. "Sorting in Experiments with Application to Social Preferences." *American Economic Journal: Applied Economics* 4(1):136–63.

[155] van Leeuwen, Boris, Charles Noussair, Theo Offerman, Sigrid Suetens, Matthijs van Veelen, and Jeroen van de Ven. 2018. "Predictably Angry - Facial Cues Provide a Credible Signal of Destructive Behavior". *Management Science* 64: 2973-3468.

[156] Le Quement, Mark, and Amrish Patel. 2018. "Communication as Gift-Exchange". Mimeo.

[157] Levine, David K. 1998. "Modeling Altruism and Spitefulness in Game Experiments". *Review of Economic Dynamics* 1: 593-622.

[158] Livio, Luca, and Alessandro De Chiara. 2019 (forthcoming). "Friends or Foes? Optimal Incentives for Reciprocal Agents". *Journal of Economic Behavior and Organization*.

[159] Loewenstein, George, Christopher Hsee, Elke Weber, and Ned Welch. 2001. "Risk as Feelings". *Psychological Bulletin* 127: 267-286.

[160] Loomes, Graham and Robert Sugden. 1982. "Regret Theory: An Alternative Theory of Rational Choice under Uncertainty". *Economic Journal* 92: 805-824.

[161] Loomes, Graham and Robert Sugden. 1986. "Disappointment and Dynamic Consistency in Choice under Uncertainty". *Review of Economic Studies* 53: 271-282.

[162] López-Pérez, Raúl. 2008. "Aversion to Norm-Breaking: A Model". *Games and Economic Behavior* 64: 237-267.

[163] Mailath, George, and Larry Samuelson. 2006. *Repeated Games and Reputations: Long-Run Relationships*. Oxford, UK: Oxford University Press.

[164] Mannahan, Rachel. 2019. "Self-Esteem and Rational Self-Handicapping". Unpublished.

[165] Mauss, Marcel. 1954. *The Gift: Forms and Functions of Exchange in Archaic Societies*. Glencoe, Illinois: The Free Press.

[166] Miettinen, Topi, and Sigrid Suetens. 2008. "Communication and guilt in a prisoner's dilemma". *Journal of Conflic Resolution* 52(6): 945–960.

[167] Möbius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2010. "Managing Self-Confidence: Theory and Experimental Evidence. NBER Working Paper 17014.

[168] Morrell, Alexander. 2014. "The Short Arm of Guilt: Guilt Aversion Plays Out More Across a Short Social Distance". Discussion Paper Series of the Max Planck Institute for Research on Collective Goods 2014-2019.

[169] Morris Stephen. 2001. "Political Correctness". *Journal of Political Economy* 109: 231–265.

[170] Myerson, Roger. 1986. "Multistage Games with Communication". *Econometrica* 54: 323-358.

[171] Netzer, Nick, and Armin Schmutzler. 2014. "Explaining Gift-exchange – The Limits of Good Intentions". *Journal of the European Economic Association* 12: 1586-1616.

[172] Nielsen, Carsten, and Alexander Sebald. 2017. "Simple Unawareness in Dynamic Psychological Games" *The B.E. Journal of Theoretical Economics* 17(1): 1-29.

[173] Nyborg, Karin. 2018. "Reciprocal Climate Negotiators". *Journal of Environmental Economics and Management* 92: 707-725

[174] O'Donoghue, Ted, and Matthew Rabin. 1999. "Doing it Now or Later". *American Economic Review* 89: 103-124.

[175] O'Donoghue, Ted and Charles Sprenger. 2018. "Reference-Dependent Preferences." In Douglas Bernheim Stefano DellaVigna David Laibson (eds.): *Handbook of Behavioral Economics,* Vol. 1, Elsevier.

[176] Ottaviani, Marco, and Peter Sorensen. 2006. "Reputational Cheap Talk". *The RAND Journal of Economics* 37: 155–175.

[177] Passarelli, Francesco, and Guido Tabellini. 2017. "Emotions and Political Unrest". *Journal of Political Economy* 125: 903-946.

[178] Patel, Amrish, and Alec Smith. 2019 (forthcoming). "Guilt and Participation". *Journal of Economic Behavior and Organization.*

[179] Persson, Emil. 2018. "Testing the Impact of Frustration and Anger When Responsibility is Low". *Journal of Economic Behavior and Organization* 145: 435-448.

[180] Piccione, Michele, and Ariel Rubinstein. 1997. "On the Interpretation of Decision Problems with Imperfect Recall". *Games and Economic Behavior* 20: 3-24.

[181] Potegal, Michael, Charles Spielberger, and Gerhard Stemmler. 2010. *International Handbook of Anger: Constituent and Concomitant Biological, Psychological, and Social Processes.* New York: Springer.

[182] Quiggin, J. 1994. "Regret Theory with General Choice Sets". *Journal of Risk and Uncertainty* 8: 153-65.

[183] Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics". *American Economic Review* 83: 1281-1302.

[184] Rauh, Michael T., Seccia, Giulio, 2006. Anxiety and Performance: A Learning-By-Doing Model. *International Economic Review* 47: 583-609.

[185] Regner, Tobias, and N.S. Harth. 2014. "Testing Belief-Dependent Models". Working Paper. Max Planck Institute of Economics.

[186] Rotemberg, Julio. 2005. "Customer Anger at Price Increases, Changes in the Frequency of Price Adjustment and Monetary Policy". *Journal of Monetary Economics* 52: 829-852.

[187] Rotemberg, Julio. 2011. "Fair Pricing". *Journal of the European Economic Association* 9: 952-981.

[188] Ruffle, Bradley. 1999. "Gift Giving with Emotions". *Journal of Economic Behavior and Organization* 39: 399–420.

[189] Schotter, Andrew, and Isabel Trevino. 2014. "Belief Elicitation in the Laboratory." *Annual Review of Economics* 6: 103-128.

[190] Sebald, Alexander. 2010. "Attribution and Reciprocity". *Games and Economic Behavior* 68: 339-352.

[191] Sebald, Alexander, and Nick Vikander. Forthcoming. "Optimal Firm Behavior with Consumer Social Image Concerns and Asymmetric Information". *Journal of Economic Behavior and Organization.*

[192] Selten, Reinhard. 1975. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games". *International Journal of Game Theory* 4: 25-55.

[193] Shalev, Jonathan. 2000. "Loss Aversion Equilibrium". *International Journal of Game Theory* 29(2): 269-287.

[194] Silfver, Mia. 2007. "Coping with Guilt and Shame: A Narrative Approach". *Journal of Moral Education* 36: 169-183.

[195] Smith, Alec. 2019 (forthcoming). "Lagged Beliefs and Reference-Dependent Utility". *Journal of Economic Behavior and Organization.*

[196] Smith, Eliot R., and Diane M. Mackie. 2007. *Social Psychology* (Third ed.). Hove: Psychology Press.

[197] Sohn, Jin and Wenhao Wu. 2019. "Reciprocity with Uncertainty about Others". Unpublished.

[198] Sobel, Joel. 2005. "Interdependent Preferences and Reciprocity". *Journal of Economic literature* 43: 396-440.

[199] Suppe, Frederick, ed. (1977), *The Structure of Scientific Theories*. Urbana: University of Illinois Press.

[200] Tadelis, Stephen. 2011. "The Power of Shame and the Rationality of Trust". Unpublished.

[201] Tangney, June Price. 1995. "Recent Advances in the Empirical Study of Shame and Guilt". *American Behavioral Scientist* 38: 1132-1145.

[202] Thompson, Frederick. 1952. "Equivalence of Games in Extensive Form". Research memorandum, U.S. Air Force, Rand Corporation.

[203] Trivers, Robert. 1971. "The Evolution of Reciprocal Altruism". *Quarterly Review of Biology* 46: 35-57.

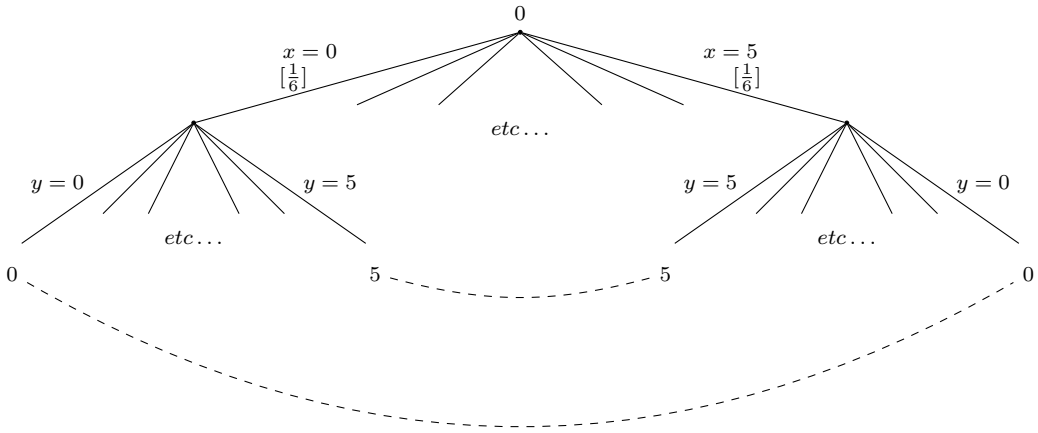
[204] van Damme, Eric, et al. 2014. "How Werner Güth's Ultimatum Game Shaped our Understanding of Social Behavior". *Journal of Economic Behavior and Organization* 108: 292-318.

[205] Woods Daniel, and Maros Servatka. 2016. "Testing Psychological Forward Induction and the Updating of Beliefs in the Lost Wallet Game". *Journal of Economic Psychology* 56: 116–125.

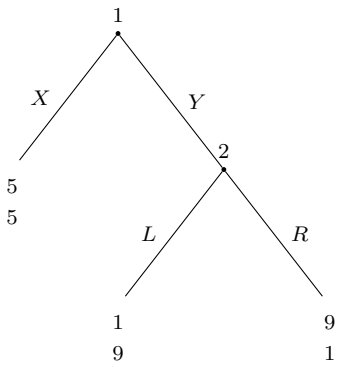
JEL game tree

April 2019

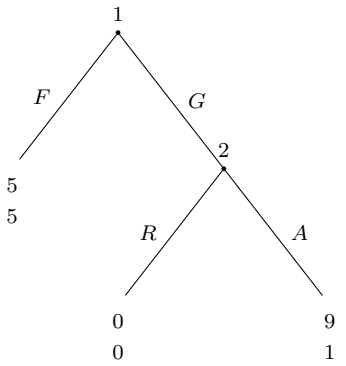
1 $G1$



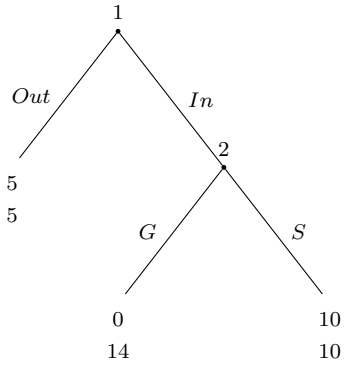
2 $G2$



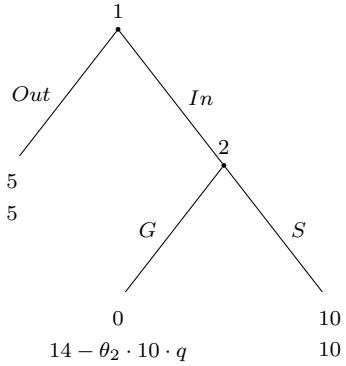
3 $G3$


1
F
G
2
5
5
R
A
0
0
9
1

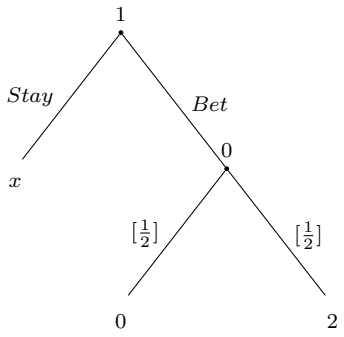
4 $G5$


1
Out
In
2
5
5
G
S
0
14
10
10

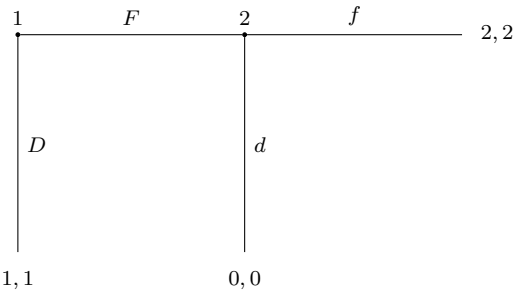
5 G_5^*


1
Out
In
2
5
5
G
S
0
$14 - \theta_2 \cdot 10 \cdot q$
10
10

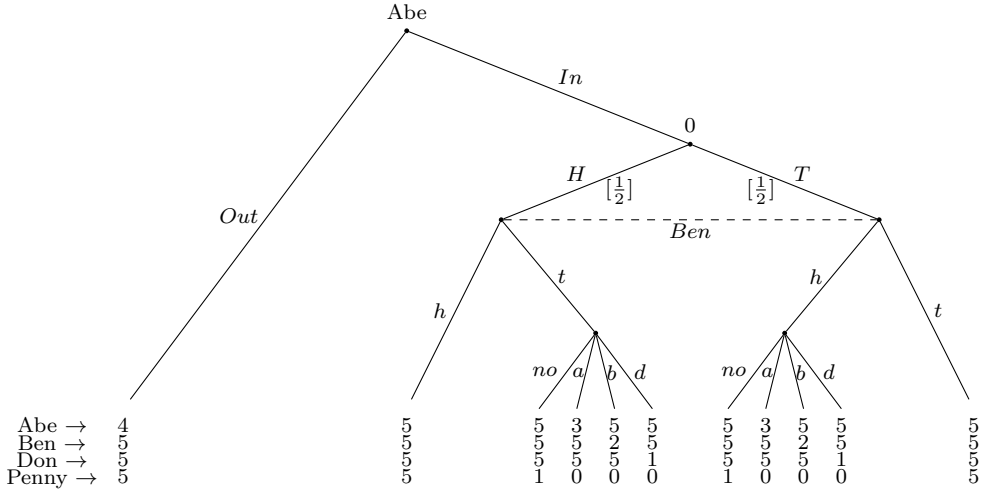
6 *G*6


1
Stay
Bet
0
x
[1/2]
[1/2]
0
2

7 *G*7


1
F
2
f
2, 2
D
d
1, 1
0, 0

8 *G*8


Abe
In
Out
0
H
[1/2]
[1/2]
T
Ben
h
t
h
t
no
a
b
d
no
a
b
d
Abe → 4 5 5 3 5 5 5 3 5 5 5
Ben → 5 5 5 5 2 5 5 5 2 5 5
Don → 5 5 5 5 1 5 5 5 1 5 5
Penny → 5 5 1 0 0 0 1 0 0 0 5

9 $G9$