Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Psychological Game Theory

Pierpaolo Battigalli & Martin Dufwenberg

Working Paper n. 646

This Version: 30 April, 2019

IGIER – Università Bocconi, Via Guglielmo Röntgen 1, 20136 Milano –Italy
hiip://www.igier.unibocconi.it

Psychological Game Theory*

Pierpaolo Battigalli & Martin Dufwenberg

April 30, 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. Consider the following categories:

- **emotions**, including guilt, disappointment, regret, frustration, anger, anxiety, shame, and fear;

*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 on (versions of) this manuscript, we thank the Editor Steven Durlauf and several referees, as well as Lina Andersson, Francesco Fabbri, Amanda Friedenberg, Senran Lin, Paola Moscariello, and Jin Sohn. Financial support of ERC (grant 324219) is gratefully acknowledged.

- **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 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 incomplete. 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).

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

[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.

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 6.

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); experimental tests (5); additional comments (6); concluding remarks (7).

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 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

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

(x, y) as $[y - x]^+ := \max\{y - x, 0\}$. Player 2 cannot observe x, but draws inferences 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, 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): $s_1(x)(y)$ 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 is strong enough ($\theta_1 > 2$), neither honesty ($s_1(x)(x) = 1$ for all x) nor selfish choice ($s_1(x)(n) = 1$ for all x) is an equilibrium. The striking implication: if $\theta_1 > 2$ then equilibrium play involves *partial* lies (in expectation).

[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) $\sum_x s_1(x)(y) > 0 \Rightarrow p_2(x'|y) = \frac{s_1(x')(y)}{\sum_x s_1(x)(y)}$. Our use of equilibrium analysis in this example does not mean that we endorse it in general; see the last part of Section 6.

To get the intuition for why this result holds, it is helpful to walk through a sketch of the proof: 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 by 2 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 & Sobel (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. No behavioral theory based on standard game theory (e.g. fixed lying costs, as modeled by Kartik 2009) could pick that up, because in such theory the information of inactive players is

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

irrelevant. PGT is needed.[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 affect the (psychological) utility of endnodes and expected utility calculations

[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 tournaments in which 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.

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** $G = \langle I, \bar{H}, \iota, p_0, (\mathcal{P}_i, \pi_i)_{i \in I} \rangle$ 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** $h = \left(a^k\right)_{k=1}^{\ell}$ (possibly including actions of the chance player) 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.
- 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

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

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 systems $\alpha_i = (\alpha_i(\cdot|h))_{h\in H\cup\mathcal{P}_i}$ 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 material payoff $\mathbb{E}[\pi_i; \alpha_i]$ (which may affect his utility *via* disappointment or

[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 by 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.

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.

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 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.

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

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

first-order beliefs profile $(\alpha_j)_{j\in I}$. Thus, we have utility functions with the general form $u_i : Z \times \left(\times_{j\in I}\Delta_j^1\right) \to \mathbb{R}$. Such utilities 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 some other 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.

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 will 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 player 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 6.)

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 $h \in H_i$, $a_i \in A_i(h)$. Subjective rationality requires that a player 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[13] and i rational plan can be non-deterministic (not a pure strategy) if and only if i is always indifferent between the pure 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 6 we briefly discuss the possible relevance of k-order beliefs with $k > 2$.

A more 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

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

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 As we mentioned above, our perspective and formal analysis differs from GP&S in important ways. Let us first address the least important one: unlike GP&S (and B&D) here 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 utility functions of the following form: $\hat{u}_i\left(a, \beta^{\varnothing}_{i,-i}\right)$, where $\beta^{\varnothing}_{i,-i} \in \Delta\left(A_{-i} \times \left(\times_{j \neq i} \Delta\left(A_{-j}\right)\right)\right)$ 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 (Section 4.2) and anticipatory feelings (4.2). Third, even if utility depends only on initial beliefs, the framework of GP&S 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 the 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., the beliefs of others.

4 Motivations

We now showcase how PGT is useful for describing many interesting forms of motivation, organizing the presentation to match the preference classes men-

tioned 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.[14] 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 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,[15] 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).[16] In G_2, if 1 believes

[14] 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.

[15] 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).

[16] This definition glosses over an important aspect that a fuller account of reciprocity would have to deal with. In some games absurd implications follow unless the calculation of "the minimum payoff j could get" is modified to not consider so-called "inefficient

there is probability p that 2 would choose L we get

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

Note that i's kindness to j has the dimension of (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.[17] 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 $p > \frac{1}{2}$.

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

strategies" that hurt both i and j. We refer to D&K (2019) for a detailed discussion of this (somewhat contentious) issue, including a response to a related critique by Isoni & Sugden (2019).

[17] 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.

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.[18] 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 manifest 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.

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 chose G would not be zero.[19]

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." Jiang & Wu (2019) discuss alternative belief-updating rules to those of D&K. Dufwenberg, Smith & Van Essen (2013) (DS&VE) modify D&K to focus on "vengeance;" players reciprocate negative but not positive kindness (achieved by replacing $\kappa_{ji}(\cdot)$ in

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

[19] 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.

(2) by $[\kappa_{ji}(\cdot)]^-$). D&K, Sebald, Wu, and DS&VE hew close to Rabin. More different approaches are proposed by Falk & Fischbacher (2006) who combine reciprocity motives with preferences for fair distributions,[20] 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 economic theory. 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, ultimatum bargaining, gift exchange, insolvency in banking, mechanism design, trade disputes, public goods, RCTs, MOUs, climate negotiations, communication, and performance-based contracts.[21]

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, in the years since PGT has been put to such use, and there is more to do. This section explains.

Guilt Psychologists Baumeister, Stillwell & Heatherton's (1994) 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 maintain and strengthen

[20] 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) and Gul & Pesendorfer (2016) present reciprocity-related ideas which are not kindness-based and do not use PGT.

[21] 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 & Nicolo (2019), Aldashev, Kirchsteiger & Sebald (2017), Nyborg (2018), Le Quement & Patel (2018), and Livio & De Chiara (2019).

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 her driver. 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.[22]

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) has to "*wait*." Choice t pins down a strategy profile (t, *wait*), 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.[23] 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) have been experimental, a few applied theory papers have explored how guilt influences marriage & divorce, corruption, cheating, framing, tax evasion, public goods, embezzlement,and expert advice.[24]

We describe B&D's (2009) notion of "simple guilt," which player i experiences when the payoff j gets ($\pi_j(\cdot)$) is lower than the payoff j initially expected ($\mathbb{E}[\pi_j; \alpha_j]$).[25] Specifically, $i \neq j$ maximizes (the expectation of) a

[22] 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.

[23] 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.

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

[25] 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.

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.[26] 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.[27]

$$[G_5 \text{ and } G_5^*]$$

Several PGT-related observations are pertinent:

(i) G_5^* is a p-game, because of the presence of belief-feature q in (3).

(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.[28]

(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 2 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.[29]

(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.

(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

[26] 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).

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

[28] 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).

[29] 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.

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).

Three further observations compare simple guilt and reciprocity:

(vi) 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.[30]

(vii) 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 one must describe and re-evaluate each player's kindness at each history.[31]

(viii) 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.[32] 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

[30] On this, see Attanasi, Battigalli & Nagel (2013).

[31] 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.

[32] 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, tends to focus on the impact of guilt when it actually occurs. The quote with which we opened this section, from Baumesister *et al.*, is exceptional in that regard.

> 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), and more recent approaches by Shalev (2000) and K&R are also related.[33] Disappointment may help explain consumption, risk-references, and savings (K&R), as well as labor supply (Abeler, Falk, Goette & Huffman 2011) and behavior on tournaments for "promotions; bonuses; professional partner-ships; elected positions; social status; and sporting trophies" (Gill & Prowse 2011, p. 495) (G&P). Most of this work (not Shalev & 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.The factor $[\mathbb{E}[\pi_j; \alpha_j] - \pi_j(\cdot)]^+$, seen in eq. (3),captures j's disappointment,[34] although in (3) it was used in 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. Below, we consider a case with $k > 0$ to make a technical point.

While (4) is applicable to any game form, and hence can shape strategic interaction generally, the clearest way to highlight the essence of disappointment is to use a one-player game. We will work with game form G_6:

$$[G_6]$$

Utility (4) looks deceptively similar to (3) but is crucially different in that i's utility depends (in part) on his plan. Such "own-plan dependence," where i's beliefs about his choices impacts the utility of his choices,can lead to subtle

[33] Shalev's and K&R's goal is not to model disappointment, but rather to tie in with Kahneman & Tversky's (1979) (K&T) classic 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.

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

complications. Our observations to follow highlight that (and see BC&D for more):

(i) Can *Stay* be an optimal plan for 1 in G_6^*? This requires

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

Similarly, *Bet* is an optimal plan if

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

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 optimal plans could never happen without own-plan dependent utility. In the standard case multiplicity of optimal plans is possible only if there is indifference, hence no loss from deviating to other optimal plans.

(ii) An interesting variation arises if $k>0$. Assume that $k=1-x$; this simplifies calculations. Eq. (5) still applies while (6) can now be re-written:

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

Could it be that neither *Stay* nor *Bet* is optimal? If so, then neither (4) nor (6) holds, which requires

$$\frac{2-\theta_1+\theta_1\cdot k}{2} < x < \frac{2+\theta_1\cdot k}{2+\theta_1}. \quad (7)$$

It is straightforward to verify that all these inequalities hold (as well as $k=1-x$) if $\theta_1 \in (0,2]$, $k\in(0,1)$, and $x \in (\frac{2-\theta_1+\theta_1 k}{2}, \frac{2+\theta_1-\theta_1 k}{2+2\theta_1})$. That is, for this parameter region rational planning rules out that 1 uses a degenerate plan. Again, this could never happen without own-plan dependent utility.

(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 limitour discussion of frustration toits role in that context, which, we argue, suggests a difference in how to model frustration and disappointment. All of this will be discussed in the next section, subsumed under the heading of...

Anger Frustration breeds anger and aggression toward others. This can have profound economic impact, though few economists studied the topic. Battigalli, Dufwenberg & Smith (2015) (BD&S) propose a broadly applicable model. While they do not develop applications, they mention pricing, domestic violence, riots, recessions, contracting, arbitration, terrorism, road rage, support for populist political candidates, and bank bail-outs as potentially interesting ones.[35] We sketch key features of BD&S's approach, starting with an example of theirs – G_7 – which is 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 was 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 diagnose the issue such that when a player evaluates her frustration, she should concentrate on what happened and what she can best achieve,

[35] 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.

going forwards, in material terms. Maybe she will be frustrated and so end up meting out a costly punishment, but if so 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 $\mathrm{F}_2(F;\alpha_i) = [(1-p)\cdot 1 + p\cdot q\cdot 2 - 2]^+ = 0$. In BD&S' model, 0 frustration breeds 0 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 to BD&S's theory. 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 then develop three models based on different notions of blame:

(i) *Simple anger*: all co-players are blamed independently of how they have chosen.[36] 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.

[36] 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.

(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 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 pa-

pers on regret in auctions.[37] 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 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: Too 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. What would 1's regret-adjusted utility be if he chose X? The answer depends on p, the probability with which 1 believes

[37] See e.g. Engelbrecht-Wiggans (1989), Engelbrecht-Wiggans & Katok (2008), Filiz-Ozbay & Ozbay (2007).

that 2 would choose L had 1 chosen Y.

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 the 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. Using our notation, they analyze portfolio choice assuming 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, that is, the pleasure experienced immediately prior to the anticipated resolution of uncertainty. This theme is explored in depth by Ely, Frenkel, & 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, if not all, of the additional sentiments listed by Elster involve belief-dependent motivation that could be explored using PGT. However, rather than pursue these topics further 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), which may explain behavior in the Dictator Game, or why people seldom give anonymously to charities (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.[38]

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 $Z_i^B = \{(x, y) : y \neq x\}$

[38] 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.

is the set of paths where i lies, and $Z_i^G = Z \backslash Z_i^B$. In a Trust Minigame (e.g., games 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).[39] 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

$$u_i(z,\alpha) = \pi_i(z) + \theta_i\left[p_{j,i}^G(z) - p_{j,i}^B(z)\right]. \tag{1}$$

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.[40] In particular, terminal beliefs of j have the form $\alpha_j(\cdot|\mathcal{P}_j(z)) \in \mathcal{P}_j(z) \times \Theta_{-j}$. Let $\mathbf{I}_i^D(\cdot) : Z \to \{0,1\}$ 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

$$u_i(z,\alpha,\theta_i) = \pi_i(z) + \theta_i^{\mathbf{I}}\left[\mathbf{I}_i^G(z) - \mathbf{I}_i^B(z)\right] + \theta_i^{\mathbf{O}}\mathbb{E}\left[\widetilde{\theta_i^{\mathbf{I}}}|\mathcal{P}_j(z);\alpha_j\right]. \tag{2}$$

Utility functions like (2) 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

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

[40] Formally, we are considering beliefs in games with incomplete information. On p-games with incomplete information see Attanasi, Battigalli & Manzoni (2016), and BC&D. Sohn & Wu (2019) analyze games where players are uncertain about each others' reciprocity sensitivities.

that 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.[41]

This is a good point to reflect on a more general theme: Can and should we study the aforementioned psychological phenomena using standard game theory (GT)? Indeed such p-games as those described above 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 (2) is replaced by $a_j \geq 0$ and 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 (2). 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 its 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.[42] However, the possibility of such transformations has also engendered the claim that PGT is, after all, not needed: choosing different "weird" assumptions about utility[43] 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 is, at best, misleading. Furthermore, nobody has shown that all interesting forms of p-games can be turned into "equivalent" standard games.[44]

[41] The signaling element may also be present in p-games with utility like (1): in the warm-up example of Section 2 report y is a(n imperfect) signal about the die roll x.

[42] 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.

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

[44] Considering claims made at seminars attended of presented by us we suspect that this is not for lack of trying.

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

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

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.

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? On the one hand, 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.

5 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.[45]

[45] 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

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 was built around that insight. Dufwenberg & Gneezy (2000) (D&G) considered versions of G_4 (recall: player 2 chooses $t \in \{0, ..., W\}$) 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* 1*'s guess* (again with rewards for accuracy).[46] 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. We make several related observations:

(i) There is a large follow-up literature exploring guilt in trust games (and usually binary trust games like G_5 rather than lost wallet games). See Cartwright (2019) for a survey.

(ii) There are also studies that rely on similar techniques for studying other forms of motivation than guilt. The pioneer to do this for reciprocity theory is Dhaene & Bouckaert (2010). And several recent studies testing aspects of BD&S' account of frustration and anger do it – see Aina, Battigalli & Gamba (2018) and Dufwenberg, Li & Smith (2018*a*,*b*), and Persson (2018).

(iii) There are issues regarding how to best measure subjects beliefs.[47] Different papers (e.g. those cited by Cartwright) take different approaches and some discuss pros & cons.

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 choice. This would imply that a subject's choice drives his SOB, rather than

what data they needed for that purpose (including eliciting particular conditional beliefs).

[46] This description is precise as regards G_4. In D&G's lost wallet games 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.at*.

[47] 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).

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.[48]

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 conclusion 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 test 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,

[48] This line of criticism has made EJT&T's approach controversial, and yet the technique has come to been frequently relied on. See e.g. Attanasi, Battigalli & Nagel (2013), Khalmetski, Ockenfels & Werner (2015), Bellemare, Sebald & Suetens (2017), Attanasi, Battigalli, Manzoni & Nagel (2019), and Dhami, Wei & al-Nowaihi (2019) all of whom critically discuss the issue.

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.

Other forms of data It may be useful to consider other kinds of data than choices and elicited beliefs to test PGT-based hypothesis. For example brain imaging data (e.g. fMRI), emotion self-reports ("please rate how strongly you feel emotion X on a scale..."), or face-reader data may be useful. Chang, Smith, Dufwenberg & Sanfey (2011) (CSD&S) pioneered to 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 (vii) in the guilt part of Section 4.2 above).[49] 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 result when motivating their own theory.

6 Additional comments

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 suspense to anxiety. Another example involves an opposite to guilt. Khalmetski, Ockenfels & Werner (2015) (KO&W) consider that i enjoys surprising j so that j gets a higher material payoff than j expected. This can be modeled by substituting $[\cdot]^-$ for $[\cdot]^+$ in (4).

We did not give either of those two sentiments their own heading in Section 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.[50] As regards enjoying surprising others, is that an "emotion"? The sentiment may be empirically relevant, but we felt it did not obviously fit any category of Section Section 4.2 so we didn't present it there.

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 KO&W's. GP&S's example does not require that the co-player is surprised in terms of material payoff. 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

[49] 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."

[50] 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).

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 eq.s (3) & (4)). Second, calculate 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 he knew he were at z. Finally, i suffers from guilt-from-blame in proportion to j's blame, and

[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.)

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

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

[53] C&D's reason is conceptual; from a contract-theoretic point G_9 may be seen to incorproate 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.

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) 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.

Social norms "A clear definition of a social norm is provided by Fehr & Schurtenberger (2018), namely 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 (see also Elster 1989, Bicchieri 2005)." This is a quote from Adda, Dufwenberg, Passarelli & Tabellini (2019) (ADP&T), who 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 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

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

work to do in this arena, and we suggest that it should involve (some possibly extended version of) PGT.[57]

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. Too 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).[58] For example, an agent with imperfect recall may store and recall, possibly at a cost, the flow of information

[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 nature of the preference involved? Is the cost of breaking a norm dependent on whether and how many others do so?

[58] 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).

he receives, thus manipulating what he is able to remember and his beliefs.[59] 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.

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 rare cases this is justified,[60] but in general the equilibrium presumption is unjustified. 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, the appropriate solution is (some version of) self-confirming equilibrium, in which beliefs may be incorrect, although consistent with evidence. In neither case is the most commonly applied solution—sequential equilibrium—generally implied.[61]

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 notions of equilibrium, we made a few related references when recalling such work.

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

[60] For example, in D&D's model, presented in Section 2, given the interpretation that player 2 is player 1's "imagined" audience (as hinted at), this may be the case; 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).

[61] 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.

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

7 Concluding remark

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.[62]

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

References

[1] Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. "Reference Points and Effort Provision". *American Economic Review* 101: 470-492.

[2] Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019 (forthcoming). "Preferences for Truth-Telling". *Econometrica*.

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

[62] See e.g. Fehr & Schmidt (1999), Bolton & Ockenfels (2000), Charness & Rabin (2002) for models, and Cooper & Kagel (2009) for a survey in that spirit.

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

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

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

[7] 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.

[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 & 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 & Organization*.

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

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

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

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

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

[17] 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.

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

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

[20] Battigalli, Pierpaolo, Martin Dufwenberg, and Alec Smith. 2015. "Frustration and Anger in Games.". Bocconi University IGIER working paper 539.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[40] 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.

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

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

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

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

[45] 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.

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

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

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

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

[50] 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.

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

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

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

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

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

[56] 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.

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

[58] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[80] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[114] 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.

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

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

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

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

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

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

[121] López-Pérez, Raúl. 2008. "Aversion to norm-breaking: A model". *Games and Economic Behavior* 64: 237-267.

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

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

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

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

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

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

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

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

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

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

[132] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

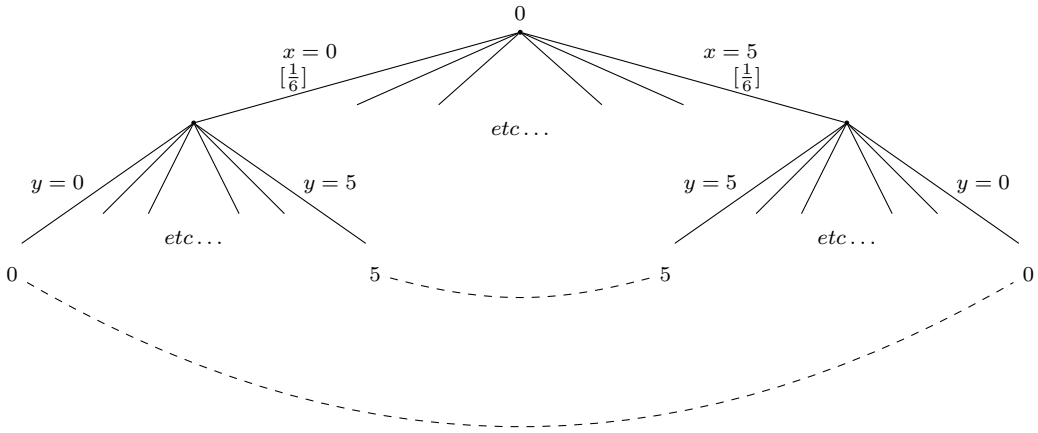
[146] van Damme, Eric, et al. 2014. "How Werner Güth's Ultimatum Game Shaped our Understanding of Social Behavior". *Journal of Economic Behavior & Organization* 108: 292-318.

JEL game tree

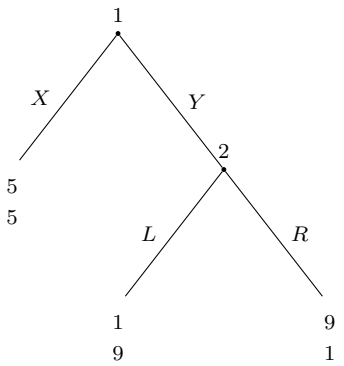
wensente9682

April 2019

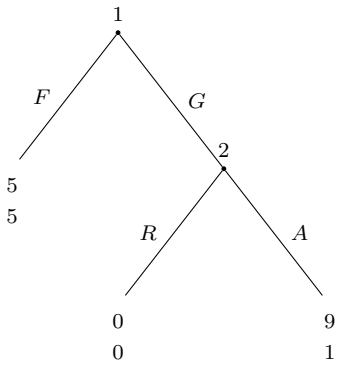
1 $G1$



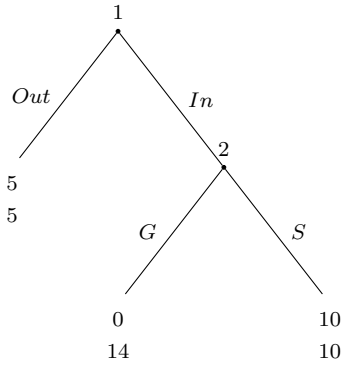
2 $G2$



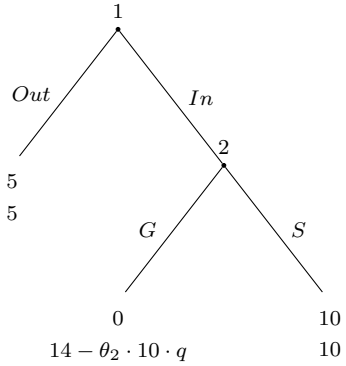
3 $G3$


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

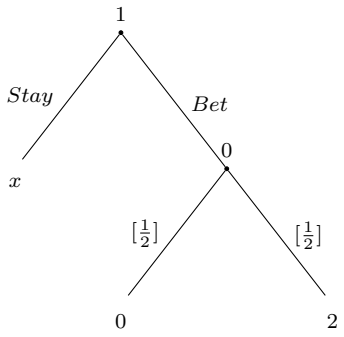
4 $G5$


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

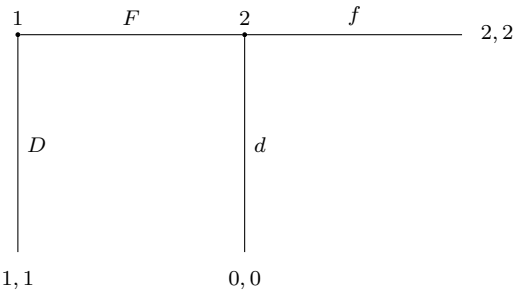
5 G_5^*


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

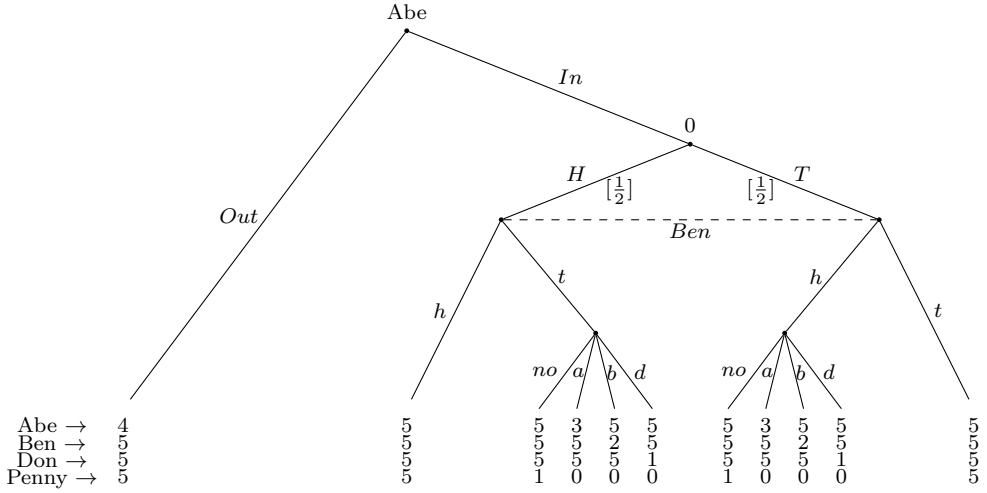
6 *G*6


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 $G9$