Belief-Dependent Motivations and Psychological Game Theory*

Pierpaolo Battigalli & Martin Dufwenberg

September 20, 2020

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 incorporating, e.g., emotions, reciprocity, image concerns, and self-esteem in economic analysis. We explain how and why, discussing basic theory, experiments, applied work, and methodology.

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

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

1 Introduction

Economists increasingly argue that a rich variety of human motivations shape outcomes in important ways. Some categories (e.g., profit-maximization, altruism, inequity aversion, maximin preferences, or warm glow) can be handled using standard tools, most notably traditional game theory. However, many other important sentiments which involve what we will call "belief-dependent motivation" defy standard analysis. A broader mathematical

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

framework called "psychological game theory" (PGT), pioneered by Geanakoplos, Pearce & Stacchetti (1989) (GP&S) and further developed by Battigalli & Dufwenberg (2009) (B&D), provides adequate modeling tools by letting the utility of outcomes depend on endogenous beliefs.[1] We explain how and why.

Among the many belief-dependent motivations that we cover, three principal categories receive particular attention:

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

For each category, we provide a detailed discussion of main features, possible applications, and a review of the relevant literature, including related experimental tests. Here in the Introduction, we provide early exposure to key ideas via three examples that illustrate these categories:

Example 1: Guilt & tipping This example involves the emotions category. Psychologists Baumeister, Stillwell & Heatherton (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 the relationship" (see p. 245; cf. Tangney 1995). That outlook is reflected in the following vignette:

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

To model this, consider game form G_1 where Tipper (player 2) chooses tip $t \in \{0, 1, ..., M\}$ and $M > 0$ is the amount of money in her wallet. The driver (player 1) is not active, and there is no future interaction. His (material) payoff is t. Choice t thus pins

[1] See also Gilboa & Schmeidler (1988) who in another pioneering contribution on "information-dependent games" anticipated some of the themes that GP&S and B&D developed in more depth. For now dated introductions to the older PGT literature, see Attanasi & Nagel (2008) and Dufwenberg (2008).

down an end-node. Tipper's material payoff is $(M-t)$. However, her utility equals $(M-t)-\theta_2\cdot[\tau-t]^+$, where τ is 1's expectation of t (which Tipper can only guess) and $\theta_2\geq 0$ is a sensitivity parameter measuring how much Tipper suffers when she lets 1 down. (Also, $[\tau-t]^+:=\max\{\tau-t,0\}$.) In words, Tipper's utility equals own money minus a pang of guilt which is proportional to how much less the driver gets than what he expects to get. Tipper's behavior in the vignette is expected utility maximizing if $\theta_2>1$. The presence of τ in her utility makes it belief-dependent, leading to what is called a "psychological game" (p-game) here given by G_1^*. The key characteristic is that utility at an end-node depends on beliefs, in this case that 2's utility at t depends on 1's beliefs (via τ). Had we had a traditional game, utilities would be defined on end-nodes independently of beliefs and Tipper's best choice would be independent of her guess of τ.

$$[G_1 \text{ and } G_1^*]$$

Terminology Example 1 illustrates key concepts we rely on throughout: A **game form** specifies the structure of a strategic situation (the "rules of the game"): the players, how they can choose, and the material consequences (typically money) of players' actions. We reserve the term **payoffs** for material consequences. Unless players are expected payoff maximizers, payoffs do *not* represent preferences over end-nodes. These are instead given by **utility functions** (or utilities). Whereas in **traditional game theory** utilities are defined on end-nodes only, in PGT they also depend on features of beliefs about behavior, like τ in Example 1, and higher-order beliefs. Such beliefs are determined by the strategic analysis, i.e., they are endogenous. We use the term **motivations** to distinguish conceptually different parts that may affect utility. Tipper is affected by *two* motivations: own money and guilt. We use the term **belief-dependent** to describe if a motivation or utility depends on beliefs. Tipper's first motivation (own money) is not belief-dependent while the second one (guilt) is, which implies that Tipper's utility is belief-dependent. We call **psychological game**, or **p-game** for short, the model obtained by appending belief-dependent utilities to a given game form.

Example 2: Reciprocity in the battle-of-the-sexes The idea that people wish to be kind towards those they perceive to be kind, and unkind towards those deemed unkind, is age-old.[2] Early academic discussions can be found in anthropology (Mauss 1954), sociology (Gouldner 1960), social psychology (Goranson & Berkowitz 1966), biology (Trivers

[2]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.

1971), and economics where the pioneer is Akerlof (1982), who analyzed "gift-exchange" in labor markets. Akerlof had the intuition that reciprocity would imply a monotone wage-effort relationship (at least up to the level of a "fair wage"), and he posited that such a relationship exists. 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, which he developed. Our second example is taken from him.

$[G_2]$

Consider game form G_2. If the players were motivated solely by material payoffs we would have a traditional game, with two equilibria: (*opera*, *opera*) and (*boxing*, *boxing*). These strategy profiles remain equilibria in Rabin's model, where players' utilities are affected by reciprocity, but (*opera*, *boxing*)[3] is an additional equilibrium. We describe the underlying intuition: The players are "unkind" to each other, in the sense that given equilibrium expectations they minimize each other's material payoffs (to be 0 rather than 1). Each player sacrifices own material payoff in the process (getting 0 rather than $\frac{1}{2}$), but the desire to reciprocate the perceived unkindness of the co-player is strong enough to make it worthwhile.

Section 2 explains in more detail why modeling reciprocity involves PGT. The reason is that kindness depends on beliefs. Here we quote Rabin (p. 1285), who compares the (*boxing*, *boxing*) and (*opera*, *boxing*) equilibria, highlighting a non-standard aspect of his model:

> In the natural sense, both of the equilibria ... are strict: each player strictly prefers to play his strategy given the equilibrium. In the equilibrium (*boxing*, *boxing*), player 1 strictly prefers playing *boxing* to *opera*. In the equilibrium (*opera*, *boxing*) player 1 strictly prefers *opera* to *boxing*. No matter what payoffs are chosen, these statements would be contradictory if payoffs depended solely on the actions taken.

Example 3: Status & conformity Our third example illustrates an image concern as modeled by Bernheim (1994). A special case goes as follows: Agents in a population are uniformly distributed on $T = [0, 2]$, where $t \in T$ is an agent's "type" of preference for "brightness of clothing". Each agent simultaneously chooses a (garment) color $c \in T$ to wear. All agents observe these choices, and form beliefs about the type of the chooser conditional on the choice. Let t_c denote the expected type of an agent who chose c. An

[3] Which is the coordination failure with smaller material incentives to deviate.

agent of type t's utility equals $-(t-c)^2-(1-t_c)^2$. In words, he suffers quadratic losses of (i) letting his chosen color deviate from the one he favors, and of (ii) status by being *perceived* as having an expected type that deviates from 1 (the "fashion standard").

Focusing on the case where agents' types are private information, Bernheim analyzes this model as a signaling game. He looks for equilibria where the "sender" (the only active player) maximizes expected utility given the way beliefs are formed, while beliefs are formed consistently with Bayes' rule given how choices depend on types. He argues that a plausible class of equilibria involve pooling at $c=1$. Under our parameterization, such pooling can be universal, if out-of-equilibrium inferences—which cannot be pinned down by Bayes' rule—satisfy (e.g.) that $t_c \in \{0,2\}$, for all $c \neq 1$.[4]

The example is interesting to us for two reasons. First, aspect (ii), mentioned above, makes utility belief-dependent,[5] and so creates a p-game G_3^*. Second, consider a modified version of G_3^*, call it G_3^{**}, where agents' types are *observed* ex post. Obviously, an agent of type t will be believed to have type t, regardless of his choice c. In the unique equilibrium, he will rely on a dominant strategy: $c=t$, so the prediction will differ from Bernheim's. A striking observation, from a game-theoretic point of view, is that the difference between G_3^* and G_3^{**} concerns information across end-nodes. It is imperfect in G_3^* but perfect in G_3^{**}. In traditional game theory, information across end-nodes never affects predictions, and is therefore not even specified. That this property does not extend to p-games shows that information at end-nodes (and more generally the information players have when they are inactive) should be carefully specified.

The preceding three examples illustrate different belief-dependent motivations that can be modeled with belief-based utility and p-games. Awareness of and interest in PGT is on the rise, yet far from universal. We explain what PGT is and what motivations can be modeled, highlighting a variety of idiosyncratic features. We discuss basic theory, experimental tests, and applied work. Although we cite a lot of papers, our primary goal is 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. This includes, in particular, the recent methodological article by Battigalli, Corrao & Dufwenberg (2019) (BC&D), a text we frequently draw connection to.

Our discussion is mainly focused on showing how to functionally represent belief-

[4] If an agent of type t stays with the proposed equilibrium he gets utility $-(t-1)^2-0$. If he deviates, the best way to do so would be to choose $c=t$, in which case he would get utility $-0-(1-0)^2$ (or $-0-(1-2)^2$), hence he cannot gain by deviating.

[5] Note that t_c is a feature of an endogenous belief, because it is derived from an initial belief about types and choices by conditioning on the observed choice.

dependent motivations. We do not critically evaluate how to best derive predictions (via "solution concepts"). Rather, we keep our analysis of strategic reasoning simple, limited to either (a few rounds of iterated) elimination of non-best replies, or to informally applying an equilibrium concept. A broader discussion of solution concepts would be an important topic, but a proper treatment warrants its own article.[6] Compared to GP&S and B&D, we greatly simplify the analysis by letting utility depend only on (own and others') beliefs about behavior and personal traits (first-order beliefs). Yet, we generalize other important aspects, e.g., by distinguishing between plans, which are beliefs/predictions about own behavior, and actual behavior. This allows us to encompass within a coherent theoretical framework essentially all the extant applied-theory models with belief-dependence,[7] including some that were not thought as connected to psychological game theory. All forms of belief-dependent motivation are thus analyzed by means of a general notion of subjectively rational planning, which accounts for the possibility of dynamically inconsistent preferences, as in—say—models of expectation-based loss aversion.

In Sections 2-5 we elucidate a wide palette of sentiments that PGT can explore, starting with the three categories of motivations mentioned above: reciprocity, which was the first application of PGT (2), emotions (3), and image concerns (4). Section 5 discusses other important, but less explored belief-dependent motivations. Section 6 builds on the models and examples analyzed earlier to delve into the abstract formal framework of PGT. Readers who like formal analysis may want to read Section 6 before the preceding ones. Section 7 discusses experiments, and Section 8 applications. Section 9 wraps up and concludes.

2 Reciprocity

Rabin's model of kindness-based reciprocity pioneered using PGT to explore the general implications of a particular motivation. He focuses on simultaneous-move game forms, as we illustrated via G_2. But, as Rabin himself points out (p. 1296)—from the perspective of applied economics—it is important to also consider extensive game forms with a non-trivial dynamic structure. Dufwenberg & Kirchsteiger (2004) took on that task,[8] and we sketch their approach. Game form G_4 (akin to their Γ_1) is useful for introducing main ideas:

$$[G_4]$$

[6] For detailed explorations of solution concepts for p-games, see B&D and BC&D.

[7] Anger from blaming intentions (Section 3.3) and guilt from blame (Section 5) are notable exceptions.

[8] The main difference between Rabin's and Dufwenberg & Kirchsteiger's approaches concerns which class of game forms is considered, but there are other differences too. See Dufwenberg & Kirchsteiger (2004, Section 5; 2019).

A crucial building block of the analysis concerns player i's kindness to j, denoted $\kappa_{ij}(\cdot)$.[9] It is the difference between the payoff (i.e., the material/monetary reward) i believes j gets (given i's choice) and a comparison payoff C that i computes as follows: C is the average of the minimum and the maximum payoff that i believe j could get, for other choices of i.[10] In G_4, if 1 believes there is probability p that 2 would choose *take*, we get

$$\kappa_{12}(stay, p) = 5 - \frac{1}{2} \cdot [5 + (p \cdot 9 + (1-p) \cdot 1)] = 2 - 4 \cdot p,$$
$$\kappa_{12}(reach, p) = p \cdot 9 + (1-p) \cdot 1 - \frac{1}{2} \cdot [5 + (p \cdot 9 + (1-p) \cdot 1)] = 4 \cdot p - 2,$$
$$\kappa_{21}(take) = 1 - \frac{1}{2} \cdot [1 + 9] = -4, \text{ and}$$
$$\kappa_{21}(give) = 9 - \frac{1}{2} \cdot [1 + 9] = 4.$$

Note that i's kindness to j has the dimension of the (expected, material) payoff of j, it ranges from negative to positive, and it may depend on i's beliefs (as it does for 1 in G_4). Player i is taken to maximize (the expectation of) a utility that depends on actions and beliefs according to a functional form of the following kind:

$$u_i(\cdot) = \pi_i(\cdot) + \theta_i \cdot \kappa_{ij}(\cdot) \cdot \kappa_{ji}(\cdot), \tag{1}$$

where $\pi_i(\cdot)$ is i's (material) payoff function and parameter $\theta_i \geq 0$ reflects i's reciprocity sensitivity. The desire to reciprocate kindness, as intuitively described in the Introduction, 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.[11] To illustrate in G_4: if θ_2 is high enough, 2 wants to "surprise" 1, i.e., 2's best reply is *take* if $p < \frac{1}{2}$ and *give* if $p > \frac{1}{2}$.

We make several PGT-related observations:

(i) Player 2 chooses between end-nodes. So, in traditional game theory, her optimal choice would be independent of beliefs. This is not the case with reciprocity. In G_4, 2's optimal choice depends on p, 1's belief. This illustrates that G_4, when played by agents motivated by reciprocity, is a p-game.

(ii) Relatedly, backward induction cannot be used to find 2's subjectively optimal choice independently of beliefs. Player 2 must consult her beliefs about p to compute his expected-utility-maximizing action.

[9] Here, and in other expressions below, the dot symbol $(\cdot)$ represents on one or more variables, such as chosen actions (terminal history reached) and beliefs.

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

[11] Player i cannot know j's beliefs and must form beliefs about $\kappa_{ji}(\cdot)$, denoted $\lambda_{iji}(\cdot)$ by Dufwenberg & Kirchsteiger, who plug $\lambda_{iji}(\cdot)$ into u_i. Our formulation, (1), conformant with Section 6 below, relies on first-order beliefs only, but has equivalent implications.

(iii) In traditional game theory, finite perfect-information games have equilibria (justifiable by backward induction) where players rely on degenerate, deterministic plans (intended choices). This is not the case in G_4, for high values of θ_2. We have not defined equilibrium here, but suppose we have a notion that requires 1 to correctly anticipate 2's plan (and that plans are carried out), and for 2 to anticipate that 1 will do so. (Dufwenberg & Kirchsteiger's equilibrium has that property.) If 2 plans to choose *take*, and 1 anticipates that 2 plans to choose *take*, then $p = 1$. But, if 1 anticipates that, then (as explained above) 2's best response would be *give*, not *take*. An analogous argument rules out an equilibrium where 2 plans to choose *give*.

Our next example, the Ultimatum Mini-game form G_5, gives further insights regarding reciprocity, and will be used for later comparisons as well:

$$[G_5]$$

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

The analysis here reflects a key feature of the 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 *fair*) and yet 2's kindness after 1 chooses *greedy* would at that time not be zero.[13]

Dealing with "bombs" The account of reciprocity theory just given glosses over a subtle issue which we now flag. To illustrate, let G_5^X be a modification of G_5 such that

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

[13] Our account has been sketchy; see van Damme et al. (2014; Section 6, by Dufwenberg & Kirchsteiger) for a fuller analysis of a large class of Ultimatum Game forms.

player 1 has a third choice at the root—X—which explodes a bomb, leaving each player with a material payoff of -100.

Recall how we defined i's kindness to j as the difference between the payoff i believes j gets and the average of the minimum and maximum payoff that i believes j could get. G_5^X can illustrate how, in some game forms, absurd implications follow unless the calculation of "the minimum payoff i believes j could get" is modified to not consider choices that hurt both i and j. In G_5 we concluded that 1's kindness when choosing *greedy* was negative ($\kappa_{12}(\textit{greedy}, p) = -2 - \frac{p}{2}$, as noted in a footnote). Reasoning analogously, in G_5^X 1's kindness of choice *greedy* would instead be positive.[14] Arguably, this is implausible. While hurting everyone would surely be unkind, not doing so should not automatically render other choices kind. The kindness (for a given p) of choice *greedy* should rather be the same in G_5^X and G_5.

Dufwenberg & Kirchsteiger (2004), as well as Rabin, propose kindness definitions that achieve this, by calculating "the minimum payoff j could get" without regard to so-called "inefficient strategies" that hurt both i and j. Their approaches, while to a degree similar in spirit, differ in details. The (somewhat contentious) issues involved are too subtle to warrant coverage here. We refer to Dufwenberg & Kirchsteiger (2019) for a detailed discussion, including a response to a related critique by Isoni & Sugden (2019).

Related literature Dufwenberg & Kirchsteiger (2004) limit attention to certain game forms without chance moves, a restriction Sebald (2010) drops, which allows him to address broader notions of "attribution" and "procedural concerns". Sohn & Wu (2020) analyze situations where players are uncertain about each other's reciprocity sensitivities. Jiang & Wu (2019) discuss alternatives to the belief-revision rules of Dufwenberg & Kirchsteiger (2004). Dufwenberg, Smith & Van Essen (2013) modify the theory to focus on "vengeance;" players reciprocate negative but not positive kindness (achieved by replacing $\kappa_{ji}(\cdot)$ in (1) by $[\kappa_{ji}(\cdot)]^-$). All these authors hew close to Rabin. Alternative approaches are proposed by Falk & Fischbacher (2006) who combine reciprocity motives with preferences for fair distributions,[15] and Çelen, Schotter & Blanco (2017) who model i's reciprocation to j based on how i would have behaved had he been in j's position.

As PGT-based models gain popularity they will be increasingly used to do applied economics. Most such work to date is based on reciprocity theory (and in particular Dufwenberg & Kirchsteiger's 2004 model). Topics explored include wage setting, voting,

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

[15] So do Rabin (1993, p. 1298) as well as Charness & Rabin (2002) in appendix-versions of their social preference models. These models and the references in the main text are PGT-based. Levine (1998), Cox, Friedman & Sadiraj (2008), and Gul & Pesendorfer (2016) present reciprocity-related ideas which are not kindness-based and do not use PGT.

framing effects, hold-up, bargaining, gift exchange, insolvency in banking, mechanism design, trade disputes, public goods, randomized control trials, memoranda of understanding, climate negotiations, communication, and performance-based contracts.[16]

3 Emotions

For a long time, neither psychologists nor economists paid much attention to emotions and how they shape behavior. We recommend Keltner & Lerner's (2010) handbook chapter which explains how while "founding figures in psychology" (in particular Charles Darwin and William James) paid significant attention to emotions, during most of the 20^{th} century and "the heyday of behaviorism ... emotions resided ... outside the purview of observable measurement" and were considered "undeserving of scientific inquiry" (p. 317).[17] Furthermore, Elster (1996, 1998) forcefully argues that economists have neglected to study the emotions, despite that the topic is potentially of great importance. In his 1996 text he goes so far as to note that "all human satisfaction comes in the form of emotional experiences" (p. 1368). He argues that by failing to recognize such an important source of utility economists are potentially failing to get a correct grip on how decisions are formed.

That view is corroborated by more recent developments in psychology. According to Keltner & Lerner, not only has (since 1980) "a robust science of emotion ... emerged" (p. 317), but it has indicated that a large variety of emotions, each one in distinct ways, impacts well-being and behavior. The causalities are complex and hardly fully understood, but a key idea that is often stressed involves what since Lerner & Keltner (2000, 2001) has been called "appraisal-tendency." Lerner, Li, Valdesolo & Kassam (2014) discuss the implications for decision making and how "appraisal tendencies are goal-directed processes through which emotions exert effects on judgments and decisions" (p. 479). The themes include how emotions affect content and depth of thought, goal activation, and interpersonal assessments.[18]

Reading these psychological discussions is highly inspiring, and we encourage economists to do so. Yet, at times, getting a full grip can be frustrating as the concepts and

[16] See Dufwenberg & Kirchsteiger (2000), Hahn (2009), Dufwenberg, Gächter & Hennig-Schmidt (2011), Dufwenberg *et al.* (2013), van Damme *et al.* (2014; Section 6), Netzer & Schmutzler (2014), Dufwenberg & Rietzke (2016), Bierbrauer & Netzer (2016), Bierbrauer, Ockenfels, Pollak & Rückert (2017) Conconi, DeRemer, Kirchsteiger, Trimarchi & Zanardi (2017), Dufwenberg & Patel (2017), Jang, Patel & Dufwenberg (2018), Kozlovskaya & Nicolò (2019), Aldashev, Kirchsteiger & Sebald (2017), Nyborg (2018), Le Quement & Patel (2018), and Livio & De Chiara (2019).

[17] Keltner & Lerner quote Skinner (1948): emotions are "the fictional causes to which we ascribe behavior" and "useless and bad for our peace of mind and our blood pressure."

[18] See also Keltner & Lerner's Table 9.3 and the related discussion of attention, certainty, control coping, pleasantness, responsibility, legitimacy, and anticipated effort.

connections tend to be, not only overwhelmingly plentiful, but also informal. We suspect and hope that some complementary clarity can be brought to the table by invoking analytical methods. PGT provides an adequate set of tools.[19] In his previous 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). While he lamented economists' dearth of attention to the issue, PGT has subsequently been put to such use, and there is more to do. In this section we focus on guilt (3.1), disappointment (3.2), anger (3.3), regret (3.4), and anticipatory feelings (3.5); we then offer some wrap-up remarks on emotions (3.6).

3.1 Guilt

Among the emotions, guilt has been explored the most using PGT.[20] Motivated by work in psychology (e.g., Baumeister *et al.* and Tangney, cited in the Introduction), Battigalli & Dufwenberg (2007) develop a model allowing exploration of how (two versions of) guilt shapes strategic interaction in a general class of game forms. While most follow-up work has been experimental (see Section 7), a few applied theory papers explored how guilt influences marriage & divorce, corruption, deception, framing, tax evasion, public goods, embezzlement, and expert advice.[21]

We provide (BC&D's account of) Battigalli & Dufwenberg's (2007) notion of "simple guilt:" Player i experiences guilt when he believes that the payoff j gets (π_j) is lower than the payoff j initially expected given j's **first-order beliefs** α_j. This expectation is denoted $\mathbb{E}[\pi_j; \alpha_j]$, and it depends on j's beliefs about (own and others') actions.[22] Specifically, $i \neq j$ maximizes (the expectation of) a utility of the form

$$u_i(z, \alpha_j) = \pi_i(z) - \theta_i \cdot [\mathbb{E}[\pi_j; \alpha_j] - \pi_j(z)]^+, \tag{2}$$

where z is the sequence of chosen actions (**terminal history**, **path,** or **end-node**). Again, $\theta_i \geq 0$ is a sensitivity parameter. As seen in the Introduction, Tipper's behavior in G_1 is

[19] See also Chang & Smith (2015) who elaborate on this theme.

[20] 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.

[21] See Dufwenberg (2002), Balafoutas (2011), Battigalli, Charness & Dufwenberg (2013), Dufwenberg & Nordblom (2018), Dufwenberg *et al.* (2011), Patel & Smith (2019), Attanasi, Rimbaud & Villeval (2019), and Khalmetski (2019).

[22] The authors 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 leads to a more complicated utility than the one seen here. However, best responses are identical, so we opt for the simpler version here.

captured if $\theta_2 > 1$. We now discuss also a trust game form G_6.[23] Assume that $\theta_1 = 0$ and $\theta_2 > 0$ to get p-game G_6^*, displayed alongside, where $\hat{\pi}_1 = \mathbb{E}[\pi_1; \alpha_1] \in [0, 10]$ denotes 1's expected payoff.[24]

$$[G_6 \text{ and } G_6^*]$$

G_6^* is a p-game, because of the presence of $\hat{\pi}_1$, an expectation derived from 1's beliefs. One may think of 2's utility as reflecting a form of "state-dependent" preference, i.e., what 2 would prefer if he knew $\hat{\pi}_1$. To maximize her utility, 2 must consult her beliefs about $\hat{\pi}_1$.[25]

In some strategic settings, powerful predictions may obtain if players reason about each other's reasoning. This may be relevant in p-games, and the emotion of guilt, as modeled in G_6^*, can illustrate this in a stark way: If $10 > 14 - \theta_2 \hat{\pi}_1$, then 2 prefers *share* over *grab*, and vice versa. No matter how high θ_2 is, if $\hat{\pi}_1$ is low enough 2 prefers *grab* over *share*. Nevertheless, 2 may reason that if 1 chose *trust* then $\hat{\pi}_1 \geq 5$, since otherwise 1 would not be rational. If $\theta_2 > \frac{4}{5}$, player 2 will then prefer *share* over *grab*, and if 1 believes that 2 will reason that way, he should choose *trust*.[26]

As argued by Charness & Dufwenberg (2006), simple guilt can explain why communication may foster trust and cooperation. Suppose G_6/G_6^* is augmented with a pre-play communication opportunity and that 2 *promises* 1 to choose *share*. If 1 believes this, and 2 believes that 1 believes this, then simple guilt makes 2 live up to her promise. A promise by 2 feeds a self-fulfilling circle of beliefs about beliefs that *share* will be chosen. Guilt, per se, does not imply such a positive effect of communication (nor does it rule out a negative effect), but it is consistent with it.[27]

Let us finally discuss the following three guilt-related distinct topics:

[23] Compare, e.g., Huang & Wu (1994), Dufwenberg (2002), Dufwenberg & Gneezy (2000), Bacharach, Guerra & Zizzo (2004), and Charness & Dufwenberg (2006).

[24] Note that $[\hat{\pi}_1 - \pi_1(trust, grab)]^+ = \hat{\pi}_1$ because $\hat{\pi}_1 \geq 0 = \pi_1(trust, grab)$.

[25] Early work on guilt (e.g., Dufwenberg 2002) plugged that second-order belief (rather than $\hat{\pi}_1$) into u_2. As explained by B&D, the two approaches are equivalent. We prefer our chosen one. The shape of 2's utility is kept simpler with only first-order belief in its domain (see Section 6).

[26] Dufwenberg (2002) calls this line of reasoning "psychological forward induction." See B&D, BC&D, and Battigalli, Corrao & Sanna (2020) for more discussion and formalization via extensive-form rationalizability.

[27] In other game forms, 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 might wish to tell j either that he had "high expectations" or that (for given expectations) the loss due to the hurtful action would be large. These are other ideas that link guilt aversion and communication, which have been explored by Cardella (2016) and Caria & Fafchamps (2019).

Counterfactual emotions In G_6^*, if 2 chooses *share* to avoid guilt, then 2 will (along the realized path) *not* experience guilt. Nevertheless, guilt has shaped the outcome. This illustrates a more general phenomenon: An emotion (it could also be, e.g., disappointment or regret, as we'll see in coming sections) need not actually realize in order to affect economic outcomes.

This observation marks a difference, to a degree, between what is the natural focus of economists and psychologists. For economists it is obvious that a counterfactual emotional experience is important, if it influences behavior and who gets what. Psychologists' discussions, by contrast, tend to focus on the impact of guilt when it actually occurs. The quote from Baumeister *et al.*, regarding guilt, which we included in the Introduction, is exceptional.

Expecting too much? Battigalli & Dufwenberg's (2007) model does not distinguish whether or not a belief by j is "reasonable," as regards whether or not guilt of i can be triggered. This assumption was made in order to keep things simple, and it could be unrealistic. For example, in G_1, if M is large and the driver expected Tipper to give away all she has then she might plausibly find the driver obnoxious, and enjoy giving nothing! Balafoutas & Fornwanger (2017) and Danilov, Khalmetski & Sliwka (2019) discuss such "limits of guilt".

Guilt vs. reciprocity With reference back to Section 2, the following points of comparison are noteworthy. First, let q denote the subjective probability assigned by player 1 to *share* in G_6; the incorporation of guilt or reciprocity has opposite connections between q and 2's preference. To see this note that the higher q the higher the payoff that 1 expects to get, and the lower the payoff 1 expects to accrue to 2, making 1 less kind toward 2. Therefore, the higher is 2's expectation of q the more (respectively, less) inclined he will be to choose *share* under simple guilt (respectively, reciprocity).[28]

Second, under simple guilt, a single utility function, that depends on initial payoff expectations and on which end-node is reached, can be applied at each history where a player moves. By contrast, to capture reciprocity motivation one must re-evaluate each player's kindness at each history.[29]

Third, recall our above remark regarding how, in G_6/G_6^*, if guilt aversion makes 2

[28] For more on this, see Attanasi, Battigalli & Nagel (2013).

[29] Herein lies *two* differences: First, a new utility function is needed for each history; see Dufwenberg & Kirchsteiger (2004) for more on this feature, which we have not illustrated very clearly since players moved once in the games we considered. Second, since kindness depends on (foregone) choice options, game-form details matter in a way that lacks counterparts with simple guilt. See BC&D for a detailed discussion of this distinction, concerning "game-form free" vs. "game-form dependent" preferences.

choose *share*, then 2 will *not* experience guilt. By contrast, if 2 were instead motivated by reciprocity, her belief-dependent motivation might be felt as she chooses *share*; at that time she perceives 1 as kind (in inverse proportion to q) which influences her utility as she chooses.

3.2 Disappointment

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

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

Belief-dependent disappointment was first modeled by Bell (1985) and Loomes & Sugden (1986). More recent work by Kőszegi & Rabin (2006, 2007, 2009) and also Shalev (2000) is technically closely related, but since it is differently motivated we write about it under the separate heading of "Belief-dependent loss aversion" in Section 5 below. Gill & Prowse (2011) argue that disappointment may help explain behavior in tournaments for "promotions; bonuses; professional partnerships; elected positions; social status; and sporting trophies" (p. 495).

Relevant needed modeling machinery was in part present already in the part on guilt of Section 3. Factor $[\mathbb{E}[\pi_j;\alpha_j]-\pi_j(z)]^+$, seen in eq. (2), captures j's disappointment, although in (2) it was used for the purpose of modeling i's guilt.[30] To let i's utility reflect disappointment we can instead look at

$$u_i(z,\alpha_i)=\pi_i(z)-\theta_i\cdot[\mathbb{E}[\pi_i;\alpha_i]-(\pi_i(z)+k)]^+, \tag{3}$$

where $k\geq 0$. In words, i's utility equals money minus a pang of disappointment which is linked to his prior expectation. Note that $k=0$ incorporates disappointment in the most straightforward way. If instead $k>0$ then disappointments are "reduced," as seen in (3).[31] Below, we consider a case with $k>0$ to make a technical point.

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

[31] Disappointment aversion may violate first-order stochastic dominance. For example, if k in eq. (3) is 0 and $\theta_i>1$, then i prefers a sure payoff $x>0$ to the lottery that yields x and $2x$ with 50% chance. The axiomatization of Gul (1991) rules this out. Cerreia-Vioglio, Dillenberger & Ortoleva (2018) derive an explicit representation of preferences à la Gul (1991).

Utility (3) looks deceptively similar to (2) but is crucially different in that i's utility depends (in part) on i's plan, that is, the part of α_i representing i's beliefs about the actions *he* is going to take (more on that in Section 6). Such "own-plan dependence," where i's beliefs about his choices impacts the utility of his choices, can lead to subtle complications as we now highlight (and see BC&D for more).

While (3) is applicable to any game form, and hence can shape strategic interaction generally, the clearest way to exhibit the essence of disappointment is to use a one-player game form with chance moves, like G_7. Assume that $0 < x < 1$ while $0 \leq k \leq \min\{x, 1-x\}$.

$$[G_7]$$

Can *stay* be a rational plan for 1 in G_7 (given (3))? This requires

$$\underbrace{x}_{\substack{\text{utility of } stay \\ \text{after planning } stay}} \geq \underbrace{\frac{1}{2} \cdot 2 - \frac{1}{2} \cdot \theta_1 \cdot [x - (0+k)]^+}_{\text{utility of } bet \text{ after planning } stay} \iff x \geq \frac{2 + \theta_1 \cdot k}{2 + \theta_1}. \tag{4}$$

Similarly, *bet* is a rational plan if

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

First, assume that $k = 0$. Inspecting (4) and (5) 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 a rational 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 rational plans could never happen without own-plan dependent utility.[32] In the standard case, multiplicity of optimal plans is possible only if there is indifference.

An interesting variation arises if $k > 0$. Could it be that neither *stay* nor *bet* is a rational plan? If so, neither (4) nor (5) holds. We would get

$$\frac{2 + \theta_1 \cdot k}{2 + \theta_1} > x > \frac{2 + \theta_1 - \theta_1 \cdot k}{2 + 2 \cdot \theta_1}. \tag{6}$$

To see that this is possible, pick a case that is easy to compute: assume that $x = k = \frac{1}{2}$, and study (6) as θ_1 increases. The leftmost term exceeds $\frac{1}{2}$ for any $\theta_1 \geq 0$, while the rightmost term is lower than $\frac{1}{2}$ for high enough θ_1 (it decreases from 1 to $\frac{1}{4}$ as θ_1 goes from 0 to infinity). All in all, for a high enough value of θ_1, (6) must hold.

We round up with two more remarks:

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

Elation This emotion, discussed by Bell (1985) and Loomes & Sugden (1986), is a sort of opposite of disappointment. It can be modeled by substituting $[\cdot]^-$ for $[\cdot]^+$ in (3) which then leads to p-games.[33]

Reference point Bell (1985) and Loomes & Sugden (1986) differ from us in the way they define rational choice/planning: We assume that realized payoff is compared to the *ex ante* (before choice) expected payoff, which depends on the agent's pre-determined plan. They instead assume that the term of comparison (reference point) depends on the actual (irreversible) choice. Since this relates to how Kőszegi & Rabin (2006, 2007, 2009) model belief-dependent loss aversion, we postpone the discussion to Section 5.

3.3 Frustration & anger

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

Anger and aggression can have profound economic impact, though few economists studied the topic. Battigalli, Dufwenberg & Smith (2019) propose a broadly applicable model. They do not develop applications, but mention pricing, domestic violence, riots, recessions, contracting, arbitration, terrorism, road rage, support for populist politicians, and bank bail-outs as potentially interesting ones.[34] We sketch key features of the approach, and start with an example from the authors—G_8—designed to make a technical point about frustration and how it compares with disappointment:

$$[G_8]$$

Suppose that if 2 is frustrated she will consider 1 an attractive target of aggression. What would she do if 1 chooses *forward*? The answer may seem intuitively obvious, but consider

[33] Elation is not discussed nearly as often as disappointment, and seems to be less often regarded as empirically relevant. In line with that, Gill & Prowse (2011) report results indicating "that winners are elated while losers are disappointed, and that disappointment is the stronger emotion" (p. 495).

[34] As the authors discuss, some of these topics have been analyzed by others 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.

what would happen if frustration were modeled as disappointment (more disappointment giving higher inclination to aggression). Building on eq. (3), there would be multiple optimal plans for 2, following the logic of (ii) in Section 3.2. If 2 plans to choose *havoc*, and if she believes 1 will choose *down*, then she would be disappointed after *forward*, hence choose *havoc* in order to hurt 1.

With outcome $(2, 2)$ available, this seems psychologically implausible. Battigalli, Dufwenberg & Smith resolve the issue by requiring players to focus on what has happened and what they can achieve in the future.

Maybe she will be frustrated and end up meting out a costly punishment, but that should be a reaction to, rather than a cause of, her frustration. This consideration leads to the following definition of i's frustration at history h:

$$\mathrm{F}_i(h; \alpha_i) = \left[\mathbb{E}[\pi_i; \alpha_i] - \max_{a_i \in A_i(h)} \mathbb{E}[\pi_i | (h, a_i); \alpha_i] \right]^+, \tag{7}$$

where $\mathbb{E}[\pi_i | (h, a_i); \alpha_i]$ is the material payoff that i expects to get, according to his first-order beliefs α_i, given history h and action a_i. Applied to G_8, let p be the probability 2 initially assigns to *forward* while q is the probability with which 2 plans to choose *bliss* (thus, α_2 is described by p and q). We get $\mathrm{F}_2(\textit{forward}; \alpha_2) = [(1-p) \cdot 1 + p \cdot q \cdot 2 - 2]^+ = 0$. Zero frustration breeds no aggression, so 2 will choose *bliss*.

While the frustration given by (7) differs from the disappointment-part of (3), it is still a belief-dependent expression. Moreover, at history h, frustration influences player i's objective, potentially making him angry and aggressive. We avoid going into technical details—see the article for that—and here just highlight some key themes. Number one is that one must now theorize about blame. Consider G_9 (where players payoffs are listed in alphabetical order, and Don is a dummy player):

$$[G_9]$$

Battigalli, Dufwenberg & Smith assume that a frustrated player becomes inclined to hurt those deemed blameworthy. They develop three models based on different blame notions. We indicate how they play out for Penny:

Simple anger: All co-players are blamed independently of how they have behaved.[35] In G_9, if Penny's anger sensitivity θ_{P} is high enough, she would choose d, going after Don whom she is most efficient at punishing.

[35] Some psychologists argue that frustrated people tend to be unsophisticated and inclined to blame in such a way; see Marcus-Newhall *et al.* (2000) for a discussion. 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.

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_9, with θ_{P} high, Penny would choose b, going after Ben, since Don is no longer blameworthy (he had no choice!), and Penny is more efficient at beating up Ben than Abe.

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_9, 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 actual 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.[36]

Finally, a comment about how these models apply to the Ultimatum Mini-game form, G_5. A comparison with reciprocity theory is of interest, as both approaches can explain the prevalence of *fair* offers and *rejections*. In both cases (anger and reciprocity), 2 may rationally plan to choose *reject* (if θ_2 is high enough, and, in the case of anger, if 2's initial belief that 1 will choose *fair* is strong enough). However, whereas in Dufwenberg & Kirchsteiger's theory it is possible that 1 chooses *greedy* even if he expects 2 to choose *reject* (since 1 then views 2 as unkind, and so may want to retaliate), this could never happen in (any of the versions of) Battigalli, Dufwenberg & Smith's theory. As it is developed, at the root a player cannot be frustrated and he must therefore maximize his expected material payoff.[37]

3.4 Regret

Despite Édith Piaf's assertion, regret can be a powerful feeling. To appreciate this, ask people who didn't sell stock while the coronavirus was ravaging China and Italy, but had not yet hit the US where stock prices remained close to all-time high, what they felt when the market crashed. Or, if they did sell, but at the bottom of the market, ask what they felt once the market recovered and they had not yet re-entered. Zeelenberg & Pieters (2007) discuss other examples and synthesize much evidence from psychology.[38]

[36] Indeed, blaming intentions implies a form of dependence on second-order beliefs. For more on higher-order belief dependence see Section 5.

[37] Game form G_5 also allows to illustrate another point. Like disappointment, frustration is own-plan dependent and this may lead to the non existence of a pure (deterministic) optimal plan. This is the case, for example, if 2 initially deems *fair* and *greedy* equally likely and $\theta_2 \in \left(\frac{1}{18}, \frac{2}{27}\right)$.

[38] See also Zeelenberg (1999) and Connolly & Butler (2006).

Research on regret starts with theoretical work by Bell (1982) and Loomes & Sugden (1982), who focus on pairwise choice. 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. 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 naturally called for when analyzing general game forms.[39]

Consider the following version of Quiggin's approach: Let Ω and A be (finite) sets of states (chosen by chance, or nature) and actions of the decision maker (= 1). The payoff function $\pi_1 : \Omega \times A \to \mathbb{R}$ has a finite range $C \subseteq \mathbb{R}$ of monetary consequences. Function $v_1 : C \to \mathbb{R}$ describe 1's "choiceless utility" (Loomes & Sugden'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

$$u_1(\omega, a) = v_1(\pi_1(\omega, a)) - f(\max_{a' \in A} v_1(\pi_1(\omega, a')) - v_1(\pi_1(\omega, a))), \tag{8}$$

where $f : \mathbb{R}_+ \to \mathbb{R}_+$ is strictly increasing. In words, i's overall utility involves pangs of regret that reflect i's evaluation of how much better off he could have been had he chosen differently. For our purposes it is useful to re-formulate this as a one-player game form 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 end-node (terminal history) (ω, a) is reached and revealed to 1, whose utility is computed using (8). Note that this is a traditional game, as 1's utility is uniquely determined at each end-node.

However, if one generalizes the above steps to apply to any game form, then one arrives at a p-game: To see this, fix an extensive game form, focus on player i, and try to compute his regret-adjusted utility at end-node z (and at the associated terminal information set; here is one more instance where terminal information will influence the analysis). To do that, one needs to figure out what would have happened had i chosen differently. This, in turn, depends both on what choices i's co-players actually made, and which ones they would have made 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.

[39] A handful of papers proposed ways, not based on PGT, to incorporate regret in particular games; see, e.g., Engelbrecht-Wiggans (1989), Filiz-Ozbay & Ozbay (2007), and Halpern & Pass (2014).

For example, consider G_4. Would 1 experience regret if he chose *stay*, and if so how much? The answer depends on p, the probability with which 1 believes that 2 would choose *take* had 1 chosen *reach*. Analogous remarks apply to, e.g., G_5, G_6, and G_8.

3.5 Anticipatory feelings

Uncertainty about the future can cause "anticipatory feelings" with negative or positive value felt in the present (cf. Loewenstein, Hsee, Weber, & Welch 2001). The anticipation of such feelings can drive behavior in earlier periods. Timing is essential to model this. The simplest setting for a meaningful discussion is one with two periods $t \in \{1, 2\}$ between three dates 0, 1, and 2. Each period t is between dates $t-1$ and t. Action profile a^t is selected in period t. To make the problem interesting, player i—the decision maker under consideration—has to be active in period 1 and another player (typically, chance) has to be active in period 2.

Anxiety is an anticipatory feeling with negative valence caused by uncertainty about future material outcomes (e.g., health, or consumption).[40] Huang (2020) argues that anxiety has major welfare consequences during pandemics. Drawing on earlier work by Kreps & Porteus (1978) on preferences for the temporal resolution of uncertainty, Caplin & Leahy (2001) put forward an axiomatic model of utility of "temporal lotteries" and consider specific functional forms. As one example, they analyze portfolio choice. Using our notation, they consider the following utility

$$u_i\left((a^1, a^2), \alpha_i\right) = -\left(\theta_i^V \mathbb{V}\left[\pi_i | a^1; \alpha_i\right] - \theta_i^E \mathbb{E}\left[\pi_i | a^1; \alpha_i\right]\right) + v_i^2\left(\pi_i\left(a^1, a^2\right)\right), \tag{9}$$

where $\mathbb{V}$ is the variance operator, θ_i^V and $\theta_i^E \geq 0$ are sensitivity parameters, and v_i^2 is the period-2 utility of the realized material outcome. The higher is the variance exhibited by i's beliefs about his material payoff, the lower is his utility. The 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 (2001) also briefly mention how their general theory can be adapted to model suspense, i.e., the pleasure experienced immediately prior to the anticipated resolution of uncertainty. This theme is explored in depth by Ely, Frankel, & Kamenica (2015). Finally, Caplin & Leahy (2004) draw on their (2001) theory to study interaction between, e.g., an anxious patient and his caring doctor, who decides whether or not to reveal information affecting the patient's anticipatory feelings.

[40] Future outcomes depend on own behavior (besides the behavior of others and chance). Thus, anxiety, like disappointment and anger, allows for the dependence of psychological utility on one's own plan. This calls for care in the analysis of rational planning. See Section 6.

3.6 Wrap-up remarks on emotions

We round up this section by collecting three remarks on distinct topics:

Valence and action-tendency Emotions have many characteristics, two important ones being *valence*, meaning the (material or psychological) 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 may want to choose which aspect to highlight, or abstract from. For example, Battigalli & Dufwenberg's (2007) models of guilt (cf. Section 3.1) are all about valence, abstracting away from action-tendency. This could well be restrictive; see, e.g., Silfver (2007) for a discussion of "repair behavior," which could be thought of in terms of an action-tendency of guilt. Similar remarks apply concerning the approaches to disappointment and regret presented in Sections 3.2 and 3.4. By contrast, Battigalli, Dufwenberg & Smith's models of frustration and anger are all about action-tendency, as frustration has no valence in their models. Again, this may be a restrictive abstraction. Frustration may, e.g., plausibly have similar valence as disappointment.

Is reciprocity an emotion? Judging by similarity of mathematical styles (in Sections 2 and 3), perhaps the answer could be yes. However, scholars working on reciprocity rarely describe what they model as involving emotions, and reciprocity usually does not figure in the list of emotions. We have chosen to structure our presentation accordingly.

Elster's and Keltner & Lerner's lists While we have covered several emotions, and highlighted their connections with PGT, we have not been exhaustive. Elster (1998) discusses 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. Keltner & Lerner (2010) offer another list (p. 330), which overlaps to a large degree but also adds contempt, disgust, embarrassment, contentment, enthusiasm, love, compassion, gratitude, awe, interest, amusement, and relief. We suspect that many of the additional sentiments listed here involve belief-dependent motivation that could be explored using PGT. However, rather than pursue these topics we propose that they hold promise for rewarding research to come.

4 Image concerns

Introspection and empirical and experimental evidence suggest that people are willing to give up some material payoffs to improve the opinion of others about them.[41] Evidence about deception can be explained by a trade-off between monetary payoff and a reduction of the perceived extent of cheating or lying (see the example below). Other models instead assume that agents try to signal that they have "good traits" such as being altruistic or fair (e.g., Bénabou & Tirole 2006; Andreoni & Bernheim 2009; Ellingsen & Johannesson 2008; Grossman & van der Weele 2017), which may explain behavior in the Dictator Game, or why people seldom give anonymously to charities, while they are happy to give non-anonymously (as shown by Glazer & Konrad 1996). Several other articles explore various forms of image concerns explaining, e.g., conformity, job-seeking effort, randomized survey-response, shame avoidance, peer evaluations, and pricing distortions.[42]

The aforementioned examples suggest two broad kinds of image about which people are concerned: others' (terminal) beliefs about (i) imperfectly observed *bad/good actions*, and (ii) imperfectly observed *bad/good traits*. Both are modeled by psychological utility functions. Section 4.1 shows how concerns for the beliefs of imperfectly informed observers about one's own behavior shape incentives, with a focus on the incentive to lie or cheat (4.1). Section 4.2 discusses reputational incentives due to non-instrumental concerns for what others think of one's own traits.

4.1 Opinions about bad/good actions

Play in a game form is represented by a terminal history, or path $z \in Z$ of actions taken by the players (including chance, when relevant). 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 a deception game form Z_i^B could be the set of paths where i lies; in a Trust Mini-game form (e.g., G_6 above and G_{11} below) i is the trustee and Z_i^B (resp. Z_i^G) contains the paths where he *grabs* (resp. *shares*).[43] Let j be an observer who obtains possibly imperfect information about the realized path z, and let $p_{j,i}^B(z;\alpha_j)$ (resp. $p_{j,i}^G(z;\alpha_j)$) denote the observer's ex post probability of bad (resp. good) deeds conditional on what he observed, given z and j's

[41] They may also care about their *own* opinions of themselves; we postpone a discussion of that case until Section 5, under the heading of "Self-esteem."

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

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

system of beliefs about actions α_j. An image concern related to bad/good deeds can be captured by a simple functional form like

$$u_i(z, \alpha_j) = \pi_i(z) + \theta_i \left[p^G_{j,i}(z; \alpha_j) - p^B_{j,i}(z; \alpha_j) \right]. \tag{10}$$

More generally, one can assume that intrinsic motivations—besides image concerns—also play a role (i (dis)likes good (bad) deeds as in Gneezy, Kajackaite & Sobel 2018 and Khalmetski & Sliwka 2019), or that i cares about the perceived distance from the standard rather than mere compliance. We expand on the second theme presenting the model of cheating by Dufwenberg & Dufwenberg (2018), which is a useful illustrative example.

Perceived cheating aversion A large recent literature explores humans' reluctance to lie or cheat using an experimental "die-roll paradigm" introduced by Fischbacher & Föllmi-Heusi (2013).[44] Dufwenberg & Dufwenberg (2018) propose a PGT-based account of this behavior. We draw on their work to illustrate how a concern with others' opinions regarding chosen actions can be modeled.

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.[45] Choice y, but not realization x, is observed by player 2, who is an "audience". In applications the audience might be a fellow citizen, but in the lab it 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 form is G_{10}:

$$[G_{10}]$$

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 G_{10}, these sets reflect player 2's end-of-play information.

Consider the following preference: Player 1 feels bad to the extent that player 2 believes that 1 cheats. Measure actual cheating at end-node (x, y) as $[y - x]^+$, i.e., cheating involves reporting a higher number than the roll and downward lies do not count as cheating. Player 2 cannot observe x, but draws inferences about x conditional on y. Let $\alpha_2(x'|y) \in [0, 1]$ be the probability 2 assigns to chance event $x = x'$ given report y, with $\sum_{x'} \alpha_2(x'|y) = 1$,

[44] See Abeler, Nosenzo & Raymond (2019) for a survey.

[45] That is, the monetary payoff function is $\pi_1(x, y) = y$.

so 2's expectation of 1's cheating equals $\sum_{x'} \alpha_2(x'|y)[y-x']^+$. Player 1's utility of (x,y) given α_2 is

$$u_1\left((x,y),\alpha_2\right) = y - \theta_1 \cdot \sum_{x'} \alpha_2(x'|y)[y-x']^+, \tag{11}$$

where $\theta_1 \geq 0$ measures 1's sensitivity to 2's expectation of 1's cheating. Note that (11) is independent of x. This reflects the fact 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.

Appending utility function (11) to game form G_{10}, we obtain a p-game because $\alpha_2(x'|y)$ is an endogenous belief, i.e., it has to be derived by strategic reasoning. Adopting the traditional equilibrium approach, the strategic analysis of this p-game is tractable and delivers testable predictions. In this game form, the relevant beliefs of player 1 about actions describe 1's *plan* (or behavior strategy), so $\alpha_1(y|x)$ is the probability that α_1 assigns to y after 1 observes x.[46] Dufwenberg & Dufwenberg solve for equilibria such that α_1 maximizes (11) given 2's beliefs, and $\alpha_2(x'|y)$ is computed as a conditional probability using correct initial beliefs, that is, $\alpha_1(x) = 1/6$ and $\alpha_2(y|x) = \alpha_1(y|x)$.[47] It can be shown that an equilibrium always exists. However, if 1's concern for his image is strong enough ($\theta_1 > 2$), neither honesty ($\alpha_1(x|x) = 1$ for all x) nor selfish choice ($\alpha_1(5|x) = 1$ for all x) is an equilibrium. The striking implication: if $\theta_1 > 2$ then equilibrium play involves *partial* lies (in expectation).

Walking through a sketch of the proof is helpful to get intuition for why this result holds: If honesty were expected by 2 then $\alpha_2(x|x) = 1$ 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 ($\alpha_1(5|x) = 1$ for all x) were expected then 2's expectation of 1's cheating would equal $\sum_x \frac{1}{6}[5-x]^+ = 2.5$; if $\theta_1 > 2$ player 1 could then increase his utility by deviating to $y = 0$ (so that perceived cheating $= 0$).

The analysis just conducted depends critically on the information 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 $\alpha_2(x|y) = 1$, implying that perceived and actual cheating coincide. If $\theta_1 > 1$ then 1's choices would be honest ($\alpha_1(x|x) = 1$ for all x); if $\theta_1 < 1$ then 1's choices would be selfish ($\alpha_1(5|x) = 1$ for all x). The partial-lies prediction evaporates. This illustrates a feature, reminiscent also of the earlier comparison of G_3^* and

[46] Player 1's initial beliefs about chance moves — exogenously given by the uniform distribution — are irrelevant because he chooses after observing x. However, 1 takes into account that 2 knows the chance probabilities.

[47] Formally, (i) $\alpha_1(y|x) > 0 \Rightarrow y \in \arg\max_{y'}(y' - \theta_1 \cdot \sum_{x'} \alpha_2(x'|y')[y'-x']^+)$ and (ii) $\sum_x \alpha_1(y|x) > 0 \Rightarrow \alpha_2(x'|y) = \frac{\alpha_1(y|x')}{\sum_x \alpha_1(y|x)}$.

G_3^{**}, that is unique to p-games. In traditional game theory, utilities are not affected by information across end-nodes, which therefore has no impact on the strategic analysis.[48]

4.2 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: now we have to consider systems of first-order beliefs about both (paths of) actions *and* traits. Suppose, just for the sake of simplicity, that actions are perfectly monitored *ex post*. Then, after the realization of any path of play (terminal history) z, each player j holds an endogenous conditional belief $\alpha_j(\cdot|z)$ about the traits of others θ_{-j}.[49] Intrinsic motivation of i is measured by parameter $\theta_i^{\mathbf{I}} \geq 0$, and i—besides liking material payoff and being intrinsically motivated—also cares about his **reputation**, that is, j's *ex post* estimate of $\theta_i^{\mathbf{I}}$. For example, i's psychological utility could be

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

where $\theta_i = \left(\theta_i^{\mathbf{I}}, \theta_i^{\mathbf{R}}\right)$ is i's trait vector, and $\left[\mathbf{I}_i^G(z) - \mathbf{I}_i^B(z)\right]$ denotes the net intensity of i's good deeds in path z.[50] More generally, j's *ex post* belief about $\theta_i^{\mathbf{I}}$ may be conditional on possibly imperfect information about the realized path of play. This allows for comparing non-anonymous and anonymous donations, or to consider the possibility that i has imperfect recall and is his own observer ($j = i$), as in the work Bénabou & Tirole (2002, 2006, 2011).[51]

Utility functions like (12) introduce a familiar element of signaling into the strategic analysis: even if i's intrinsic motivation to do good ($\theta_i^{\mathbf{I}}$) is low, he may be willing to pay a material cost to make j believe that $\theta_i^{\mathbf{I}}$ is high, hence that i is a "good guy". The simplest models of this kind are signaling games where only the sender is active and the receiver is a mere observer.[52]

[48] This explains why in traditional game-theoretic analysis information sets over end-nodes are usually not drawn, even if such information is objectively determined by the rules of interaction.

[49] Formally, we are considering beliefs in games with incomplete information (see Section 6).

[50] In the binary case considered at the beginning of this section, $\mathbf{I}_i^D(\cdot)$ is the indicator function of Z_i^D, the set of paths where i made good ($D = G$) or bad ($D = B$) deeds.

[51] See below. Note that Bénabou & Tirole (2006) consider a model similar to the one in the main text ($j \neq i$), but also put forward a reinterpretation with imperfect recall where j is a future self of i.

[52] Readers may fail to recognize that Bénabou & Tirole's (2006) model is a signaling game, because they choose not to frame it explicitly as such, making it seem more like a decision problem. But they indirectly hint (in footnote 17) at the fact that they are considering a refinement of signaling equilibria.

A noteworthy application of this approach concerns privacy. Depending on available technology and regulation, what we do may be monitored even when it does not affect the material payoff of anybody else. Many people seem to care about this and in Western countries there is a consensus that privacy should be protected. Gradwohl & Smorodinsky (2017) model this by considering functional forms such that (as in eq. 12), for each (z, θ_i), $u_i(z, \alpha_j, \theta_i)$ depends on the *ex post* belief of the "audience" j about θ_i. In particular they assume that—other things being equal—the agent either wants j's posterior to be the same as the prior, or dislikes being identified. Focusing on the simple case where i is the only active agent and actions are observable by j (lack of privacy), they analyze the pooling and separating (Bayesian perfect) equilibria of the resulting signaling game. Pooling distorts actions from the first best that would obtain under perfect privacy. Separation inflicts a psychological utility loss due to identification.

Another application concerns identity. According to Hupkau & Maniquet (2018), one's own personal traits are part of an agent's identity, and he can suffer from the discrepancy between his true identity and others' perceptions of it.[53] This in turn is affected by actions *via* signaling, possibly causing inefficiencies. For example, high-type agents may refrain from requesting useful service from a provider to avoid being pooled with low-type agents. Bénabou & Tirole (2011) instead consider forgetful agents who "care about who they are". Their actions depend on such self-perception, which may be forgotten later on. Thus, actions are also identity signals for the future self, introducing an identity-investment concern in the choice of the current self.[54]

5 More motivations

The previous sections focused on the three categories of motivation mentioned in the Introduction. We now complement that by discussing additional forms of belief-dependent motivations, and broader related issues.

Opposites Sometimes a meaningful belief-dependent motivation takes an "opposite" form of another sentiment, switching a sign or replacing $[\cdot]^+$ with $[\cdot]^-$ in the utility formula. We already saw examples in Sections 3.2 and 3.5, where, elation was compared to disappointment, and suspense to anxiety.

[53] In this case, the third element in eq. (12) is replaced by $\ell\left(\left|\theta_i^{\mathbf{I}} - \mathbb{E}\left[\widetilde{\theta}_i^{\mathbf{I}}|z;\alpha_j\right]\right|\right)$, where $\ell : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing (e.g., quadratic) loss function.

[54] Bénabou & Tirole's model is a p-game, because utility depends on an endogenuos belief, but—unlike Hupkau & Maniquet—they do not make the link to PGT explicit.

Another example involves an opposite to guilt. Ruffle (1997) and Khalmetski, Ockenfels & Werner (2015) consider situations in which player i enjoys surprising j, so that j gets a higher material payoff than j expected. See also Dhami, Wei & al-Nowaihi (2019). This can be modeled by substituting $[\cdot]^-$ for $[\cdot]^+$ in (2). Is surprising others this way an "emotion"? Maybe yes; obviously it is a kind of joy, which is often listed as an emotion.

The desire to surprise has venerable PGT-ancestry. GP&S explored the idea in their verbally presented opening example, although a different variety than the work cited above. GP&S's example does not require surprise in terms of material payoff.[55] Here is the quote (from p. 62), illustrating the sentiment and a feature idiosyncratic to p-games:

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

Belief-dependent loss aversion When we discussed disappointment, in Section 3, we mentioned how that sentiment is closely related to ideas explored by Kőszegi & Rabin (2006, 2007, 2009) and by Shalev (2000). The goal of these authors, however, is not to model disappointment, but rather to tie in with Kahneman & Tversky's (1979) work on prospect theory. Kőszegi & Rabin model prospect theory'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. Kőszegi & Rabin allow for losses in many dimensions, e.g., in $n+1$ dimension if there are n goods as well as money. To capture that, we would have to augment the framework in Section 6, allowing i's outcome function π_i to be vector-valued.

The key features we highlighted in regards to disappointment in Section 3.2 have counterparts in the work of Kőszegi & Rabin. Most notably the feature of own-plan dependent utility is there, and it may lead to multiplicity of non-equivalent rational plans, as well

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

as non-existence of pure rational plans. Beyond those technical similarities, details differ quite a lot. The exact way in which Kőszegi & Rabin define belief-dependent loss is different from the way that Bell (1985) and Loomes & Sugden (1986) (and we) define disappointment, and they also consider more notions of rational planning than we did when we discussed disappointment. The recent and penetrating survey on "Reference-Dependent Preferences" by O'Donoghue & Sprenger (2018) discusses all of these aspects in depth, so we refer to their text (and especially their Sections 5-7) for further details. Here we only mention one aspect. The notion of rational plan that we illustrated in Section 3.2 corresponds to Kőszegi & Rabin's concept of "personal equilibrium." They also consider (i) a refinement, "preferred personal equilibrium," that—in case of non-trivial multiplicity—selects the personal equilibrium most favorable to the initial self, and (ii) another concept, "choice-acclimating personal equilibrium," whereby the referent to which realized outcomes are compared is determined by the actual (irreversible) choice of the agent rather than his *ex ante* plan, i.e., it is the expected outcome conditional on the agent's action.

Consistently with Kőszegi & Rabin's explicit reference to different time frames, we can accommodate such distinctions in the PGT framework by explicitly introducing time periods, which in turn may comprise multiple stages (see Section 6 and BC&D): Endgame utility is the sum of the utilities of different periods, the referent for one-period gain-loss utility is determined by beginning-of-period beliefs. With this, personal equilibrium refers to rational planning in one-period decision problems, whereas choice-acclimating equilibrium applies to two-period situations where the agent chooses in the first period and uncertainty realizes in the second. Except for differences concerning the exact definition of the referent, the approach of Bell (1985) and Loomes & Sugden (1986) fits the choice-acclimating equilibrium.

As regards applied work, Kőszegi & Rabin discuss consumption, risk-preferences, and savings. O'Donoghue & Sprenger discuss papers about endowment effects, labor supply, job search, pricing, and mechanism design.

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" and "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 is imperfectly known by i. Such trait could be general intelligence, or ability. Player i's utility is increasing in his *ex post* estimate of $\theta_{0,i}$ conditional on the path of play z,[56] as in function

$$u_i\left(z,\alpha,\theta\right)=\pi_i\left(z,\theta\right)+v_i^e\left(\mathbb{E}\left[\widetilde{\theta_{0,i}}|z;\alpha_i\right]\right), \tag{13}$$

[56] Path z may include a randomly chosen output, whose distribution depends on $\theta_{0,i}$.

where the "ego-utility" v_i^e is increasing, and we allow material payoff 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 than exposing himself to the risk of discovering that his ability is low.[57]

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

A few economic studies of self-esteem model utility in line with our description here, although (unlike Mannahan) they do not make the PGT-connection explicit. See Kőszegi (2006), Eil & Rao (2011), Möbius, Niederle, Neihaus & Rosenblat (2011), Sebald & Walzl (2015), and Kőszegi, Loewenstein & Murooka (2019).[58]

Higher-order belief-dependence The framework presented in Section 6 restricts the domain of a player's utility to depend on beliefs (own and others') up to only the first order.[59] This is enough to handle almost all forms of motivation that to date have been modeled using PGT.[60] The main exception is Battigalli & Dufwenberg's (2007) model of guilt-from-blame.[61] We now indicate how that sentiment works in an example designed to provide a contrast with simple guilt (as presented in Section 3). Guilt-from-blame plugs a third-order belief into the domain of a player's utility, so we leave the framework of Section 6. We sketch the approach without going into all formal details:

> 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 (2) & (3)). Calculate how much of $[\mathbb{E}[\pi_j; \alpha_j] - \pi_j(z)]^+$ could have been averted had i chosen differently; this is how much i let j down, $LD_{ij}(\alpha_j, z)$. Second, calculate i's initial belief regarding $LD_{ij}(\alpha_j, z)$. Third,

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

[58] The aforementioned work of Bénabou & Tirole may be interpreted as modeling self-esteem, but their approach relies on imperfect recall and self-signaling.

[59] 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 calculate a best response.

[60] This includes reciprocity, if formulated as in Section 2. (As we noted in a footnote there, others use a different formulation with utilities that depend on second-order beliefs.)

[61] For other exceptions see B&D (p. 14) and Battigalli, Dufwenberg & Smith's (2019) model of anger from blaming intentions.

for each z, calculate j's terminal belief regarding i's initial belief regarding $LD_{ij}(\alpha_j, z)$; this is how much j would blame i if j knew he were at z. Finally, i suffers guilt-from-blame in proportion to j's blame, and i's utility trades off avoidance of that pang against i's material payoff.

Battigalli & Dufwenberg (2007; see Observation 1) prove that simple guilt and guilt-from-blame sometimes have similar implications. However, this is not true in general. To illustrate, consider G_{11}, a modified version of G_6 in which even if 2 chooses *share* there is a $\frac{1}{6}$ probability that 1 gets a material payoff of 0. Moreover, if 1 gets 0 then 1 is not informed of 2's choice.

$$[G_{11}]$$

What we said about simple guilt and (2) in Section 3.1 has its analog with G_{11}. We used G_6 merely because it is more spare.[62]

If player 2 is instead motivated by guilt-from-blame then the implications are different in G_6 and G_{11}. As in the previous comparison between guilt and reciprocity (Section 3.1), let q denote the subjective probability assigned by 1 to *share*. If 2 interprets the observed action *trust* as intentional (i.e., not the result of a "tremble"), then 2's updated belief about 1's expected payoff $\hat{\pi}_1$ is determined by 2's updated belief about q. In particular, both in G_6 and in G_{11}, if 2 believes that $q = 1$, then he believes that $\hat{\pi}_1 = 10$. In G_6, following *trust*, if 2's second-order beliefs assign probability 1 to $q = 1$, then for a high enough θ_2 player 2's best response is *share*. This is true just as it would be also under simple guilt. In G_{11}, however, following *trust*, if 2's second-order belief assigns probability 1 to $q = 1$, then player 2's best response is *grab* 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 *grab*. Therefore, 2 can *grab* with impunity.[63]

G_{11}, with guilt-from-blame appended to it, joins models of image concern such as G_{10} in illustrating the critical role information across end-nodes can play in p-games. Modify G_{11} such that 2's doubleton information set is broken up into two singletons. I.e., if 1 gets 0 then 1 *is* informed of 2's choice.[64] The logic of the previous paragraph no longer applies. In the modified version of G_{11} guilt-from-blame and simple guilt work similarly.

Social norms Fehr & Schurtenberger (2018, p. 458) define social norms as "commonly known standards of behavior that are based on widely shared views of how individual

[62] Charness & Dufwenberg (2006) (cited in Section 3.1) actually used G_{11} rather than G_6. Their reason is conceptual; from a contract-theoretic viewpoint G_{11} may be seen to incorporate an element of "moral hazard" which is absent in G_6. See Charness & Dufwenberg (2006, p. 1582).

[63] The logic is similar to that we illustrated in regards to G_{10} in Section 4.1.

[64] Tadelis compares behavior in experimental treatments that resemble G_{11} as well as the variation that we are describing here.

group members ought to behave in a given situation." Similar ideas are discussed by Elster (1989), Bicchieri (2006), Andrighetto, Grieco & Tummolini (2015), and Cartwright (2019).

D'Adda, Dufwenberg, Passarelli & Tabellini (2020) develop a model for a restrictive context (a form of Dictator Game) where the central notions concern a player's conception of "the right thing to do" and a proclivity to do what *others* believe is the right thing to do, especially if there is consensus about this (which would then be an ideal case of a social norm). Departing from the social norm entails an element of disappointing the expectations of others, and the authors explore the idea that decision makers are averse to doing so. In this regard, the motivation resembles guilt, as modeled in Section 3.1. However, d'Adda *et al.* consider players' expectations regarding how one *ought* to behave, rather than regarding how one will *actually* behave. This marks a way that the approach is not formally captured by p-games, as we have described them in this paper.

Many scholars wrote papers about social norms, but few proposed formal models, in particular ones that can be generally applied.[65] There is work to do, and we suggest that it should involve (some extended version of) PGT.[66]

Punishing transgressors Several motivations that we discussed incorporate some form of desire to punish those who, somehow, "misbehave." Negative reciprocity (Section 2) and anger (Section 3.3) have such features built in, and other notions may quite naturally be extended in that direction. For example, Sebald & Walzl (2015) explore the idea that player i may wish to be unkind (as in reciprocity theory) to j if j produced information (e.g., a performance review) that reduced i's self-esteem. Another example could be if someone is motivated to punish those who violate a social norm.

Recent work by Molnar, Chaudhry & Loewenstein (2020) highlights an intriguing aspect of such punishment that adds a feature of belief-dependent motivation (addressable via suitable modifications of the motivations we just mentioned). Namely, the punisher may care about whether transgressors are able to figure out that they are being sanctioned. The PGT-connection would be that the punisher's utility depends on the transgressor's belief regarding the punishers choice (and maybe also his intention).

Emotion carriers In most of the models we discussed, the belief-dependent part of a player's utility was built up with reference to particular material payoffs. For example, following Battigalli & Dufwenberg (2007), player 2's guilt in G_4 has the dimension of

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

[66] 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 (or, possibly, a set thereof)? If people like to follow norms, what exactly is the nature of the preference involved? Is the cost of breaking a norm dependent on whether and how many others do so?

(expected) material payoff of player 1. And in Battigalli, Dufwenberg & Smith's model, player i's frustration has the dimension of (expected) material payoff of i. This is *not* a necessary feature of belief-based 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 Battigalli, Dufwenberg & Smith (in their Section 5) mention how in principle frustration may depend on regret of a previous decision, unexpected perceived unfairness, or negative shocks to self-esteem.

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

There are formal models of unawareness—see Fagin & Halpern (1988) for a pioneering effort—and recent work develops related techniques for games. Heifetz, Meier & Schipper (2006) is a key contribution (and Burkhard Schipper provides an "Unawareness bibliography" with further references on his homepage at UC Davis). It is natural to imagine that belief-dependent motivation interacts with unawareness. For example, negative surprises that reveal previously unforeseen danger could instill fear; so, if 9/11 involved unawareness, then the occurrence of that event might have consequences for subsequent demand for air travel or supply of airport security.

Exploring unawareness using PGT seems potentially interesting, but we only know one paper on the topic: Nielsen & Sebald (2017). We quote the informal example they open with (pp. 2-3). It nicely illustrates how unawareness may interact with a belief-dependent motivation (namely, guilt):

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

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

Motivated beliefs The 2016 summer issue of the *Journal of Economic Perspectives* contains an interesting symposium on "Motivated Beliefs," with an introduction by Epley & Gilovich (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.[67] The idea is this: Beliefs affect people's well-being. This, in turn, affects how they reason, control information, and gather & evaluate evidence. To some extent, it is argued, they may even *choose* their beliefs, although such choice may be unconscious and constrained by reality-checks and various costs of having faulty beliefs. Epley & Gilovich 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.

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 formal, and PGT may provide relevant tools for scholars 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).[68] For example, an agent with imperfect recall may store and recall, possibly at a cost, the flow of information he receives, thus manipulating what he is able to remember and his beliefs.[69]

[67] See also Loewenstein & Molnar (2018) who discuss related themes.

[68] Indeed, we touch on examples of this sort, e.g. in Section 5 on self-esteem, and also where we discussed the impact of different information structures (Section 4.1 and the part on "higher-order belief-dependence" of this section).

[69] Compare with Bénabou & Tirole (2002, p. 871) and their citation from Darwin (1898), where the great

As regards addressing other cognitive phenomena that the literature on motivated beliefs has discussed (e.g., modes of reasoning, evaluating evidence, and unconscious manipulation of beliefs) it seems less clear how PGT may provide useful tools. However, we are optimists and conjecture that PGT might prove useful for approaching those topics as well.

6 Formal framework

The previous exposition of PGT relied on intuition, examples, and some formulas involving game forms and beliefs that could be understood in the context of specific models. A better appreciation of PGT requires an understanding of its formal framework, which is necessarily abstract. In this section we exhibit details, focusing on a restricted class of game forms that covers all the examples of this paper except G_3^* and G_3^{**}. We consider p-games obtained from *finite multistage* game forms 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.[70] We allow for the possibility of imperfect terminal information, which—as highlighted in the previous sections—may matter for psychological reasons.[71]

The key feature of the analysis is the representation of players' beliefs about how the game form is played (first-order beliefs), and their beliefs about beliefs (second-, or higher-order beliefs), as such beliefs affect the (psychological) utility of end-nodes and expected utility calculations at non-terminal nodes. To simplify the analysis, we mostly assume common knowledge of the rules of the game form and of players' utility functions, i.e., *complete information*. Yet, as illustrated in Sections 4 and 5, incomplete information has to be addressed when we analyze specific motivations such as image concerns and self-esteem, where utility depends on terminal beliefs about unknown personal traits. Furthermore, incomplete information is pervasive in experiments and in the field, therefore we later hint at how to generalize the analysis to take this into account.

Our *conceptual perspective* mostly relies on B&D, rather than GP&S. The reason is that GP&S only encompasses utilities that depend on players' *initial* hierarchical beliefs,

scientist describes how he manipulates memory of unpleasant facts to counteract unconscious removal.

[70] G_3^* and G_3^{**} involve game forms with a continuum of actions. G_9 is covered only if modified (or interpreted) such that chance and Ben move simultaneously.

[71] 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 (3.5). Our analysis works "as is" under the assumption that non-active players have the coarsest information consistent with perfect recall. Second, we assume that material 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.

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 As in our earlier work on PGT, we adopt a representation where the primitive elements are players' actions, rather than the older, abstract formalism where the primitives are the nodes in a tree; nodes are derived elements. We start with a **game form** $G = \left\langle I, \bar{H}, \iota, p_0, (\mathcal{P}_i, \pi_i)_{i \in I} \right\rangle$ with the following elements:

- I is the set of **players** not including **chance**, who is player 0; the set of personal players plus chance is $I_0 = I \cup \{0\}$.

- $\bar{H}$ is a finite set of possible sequences of action profiles, or **histories** $h = \left(a^k\right)_{k=1}^{\ell}$ (for different values of the length ℓ, and possibly including actions of chance) with a *tree*-structure: every prefix of a sequence in $\bar{H}$ (including the empty sequence $\varnothing$) belongs to $\bar{H}$ as well. Thus, histories in $\bar{H}$ correspond to nodes of the game tree and $\varnothing$ is the root. Set $\bar{H}$ is partitioned into the set of **non-terminal** histories/nodes H and **terminal** histories (paths, end-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 histories where i is active, and the set of feasible action profiles is

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

  with $A_i(h)$ denoting 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 i's **material payoff** function.

To illustrate, in the reporting game form G_{10} (Section 4.1), $I = \{1, 2\}$ where 1 is the only active player and 2 is an observer whose payoff we can ignore, $\bar{H} = \{\varnothing\} \cup \{0, ..., 5\} \cup Z$ with $Z = \{0, ..., 5\}^2$, $\iota(\varnothing) = \{0\}$, $p_0(x|\varnothing) = \frac{1}{6}$ (initial die roll), $\iota(x) = \{1\}$, $\pi_1(x, y) = y$, $\mathcal{P}_1(x, y) = \{(x, y)\}$, and $\mathcal{P}_2(x, y) = \{0, ..., 5\} \times \{y\}$ for every $(x, y) \in Z$ (2 only observes 1's report y).

Beliefs We model the **first-order beliefs** of (personal) player i as a system $\alpha_i = (\alpha_i(\cdot|h))_{h \in H \cup \mathcal{P}_i}$ of conditional probabilities about paths of play $z \in Z$. We are not assuming that i observes h when he 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: (i) α_i is consistent with p_0, (ii) the chain rule holds, and (iii) i's beliefs about simultaneous or past and unobserved actions of other players do not depend on i's chosen action.[72] 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*. For example, i's initially expected material payoff $\mathbb{E}[\pi_i; \alpha_i]$ (which may affect his utility *via* disappointment or frustration) depends on both $\alpha_{i,i}$ and $\alpha_{i,-i}$. As we further explain below, the interpretation is that i plans his contingent choices given his conjecture and thus ends up with an overall system of beliefs about paths.

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

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

[72] For example, consider a variation of G_{10} 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.

We point out two conceptually relevant differences with B&D: (i) 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. (ii) 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 beliefs about the behavior of *everybody*.[73] 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's plan was indeed $\alpha_{j,j}$ but that he took an unplanned action by mistake (a kind of "tremble" as in Selten 1975). The analysis of B&D instead rules this out: in B&D every observed action is *necessarily* interpreted as a planned choice (cf. our discussion of guilt aversion in Section 3.1). In sum, our framework is sufficiently expressive to model players' intentions, their perceptions of the intentions of others, and how such perceptions are affected by observing actions. This is important in standard games to elucidate the difference between, say, forward- and backward-induction reasoning. It is even more important when players care intrinsically for the intentions of others, as with many forms of belief-dependent preferences.[74]

Traditional utility Before we describe the belief-based utilities that are characteristic of p-games, it may be helpful to recall how utilities are defined in traditional game theory. Namely, player i's utility has the general form $u_i : Z \to \mathbb{R}$. This does not imply that i is "selfish". Caring only about own material reward is just a special case ($u_i(z) = \pi_i(z)$ for all $z \in Z$), but i could alternatively be motivated by a host of "social preferences" including altruism, inequity aversion, maximin preferences, or warm glow.[75] However, the forms of motivation that we discussed are ruled out as they require a richer notion of utility.

Psychological utility and p-games As argued by BC&D, most forms of belief-dependent motivations for a given player i can be modeled by assuming that, for some terminal history z, i's utility for reaching z depends on the first-order beliefs profile $(\alpha_j)_{j \in I}$. Thus, we have utility functions with the general form $u_i : Z \times \left(\times_{j \in I} \Delta_j^1\right) \to \mathbb{R}$. These typically involve

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

[74] See BC&D and Battigalli, Corrao & Sanna (2020).

[75] For prominent examples of specific functional forms, see, e.g., Fehr & Schmidt (1999), Bolton & Ockenfels (2000), or (the main text model of) Charness & Rabin (2002).

both the material payoffs and some features of own or others' initial, interim, or terminal first-order beliefs. For example, in the cheating model of Section 4.1, 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 (material) payoff and his initial belief (cf. Section 3.1 on guilt aversion). In both cases, i's utility of terminal histories depends on payoffs and the (unknown) system of first-order beliefs of another player. This is like a standard state-dependent utility function: utility does not only depend on the outcome (determined by z), but also on some aspect which is unknown to the agent, which in our case is the initial (first-order) beliefs of others, as well as the rules according to which others change their beliefs as the play unfolds. As noted by B&D, in this case the maximization of the expected value of u_i can be analyzed with standard techniques, leveraging on the dynamic consistency of subjective expected utility maximizers.

For other motivations like aversion to disappointment (Section 3.2), or belief-dependent loss aversion (Section 5), 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 showed in Section 3.2 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 emotions like frustration and anger (Section 3.3) and to anticipatory feelings with negative or positive valence like anxiety, or suspense (Section 3.5). Essentially, i's plan $\alpha_{i,i}$ must form an "intra-personal equilibrium" given his overall belief system (α_i, β_i) (cf., e.g., Kőszegi 2010 and the relevant references therein). Next, we explain this in detail.

The combination of a game form and psychological utilities for all players gives a p-game. We consider only p-games where the belief-dependence of utility is limited to first-order beliefs. The part of Section 5 on "higher-order belief-dependence" mentions exceptions.

Subjective rationality 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. $h \in H_i$. The concatenation of h with an action profile $a \in A(h)$ is a history $h' = (h, a) \in \bar{H}$. With this, each action $a_i \in A_i(h)$ yields expected utility

$$\bar{u}_{i,h}(a_i; \beta_i) = \sum_{a_{-i} \in \times_{j \in \iota(h) \setminus \{i\}} A_j(h)} \alpha_{i,-i}(a_{-i}|h)\, \mathbb{E}[u_i|\,(h, (a_i, a_{-i}))\,; \beta_i]. \tag{14}$$

Belief system (α_i, β_i) satisfies **rational planning** if the following incentive-compatibility condition holds: every action that i expects to take with positive probability is a local best reply, that is,

$$\alpha_{i,i}(a_i|h) > 0 \Rightarrow a_i \in \arg\max_{a_i' \in A_i(h)} \bar{u}_{i,h}(a_i'; \beta_i) \tag{15}$$

for all $h \in H_i$ and $a_i \in A_i(h)$. **Subjective rationality** *requires that player i plans rationally given his (second-order) beliefs and carries out his plan when given the opportunity.* Thus, consistency between plan and behavior is not a necessity. Indeed, a player's plan of how to play cannot constrain his behavior, it can only guide it. Consistency between plan and behavior should be a rationality condition. This is the perspective adopted here and in more recent work on p-games.[76]

The catch in equations system (15) is that—as explained above—the plan $\alpha_{i,i}$ is part of the overall belief system β_i; thus, rational planning is a kind of fixed-point condition. 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 can be obtained with a "folding-back" computation and is equivalent to the standard sequential rationality condition.[77] With this, i's 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 $\alpha_{i,i}$,[78] first, it may be impossible to satisfy the standard sequential rationality condition, second, deterministic rational plans may not exist. The reason is that rational planning (15) is an "intrapersonal-equilibrium" condition. Similar to how in traditional n-person games pure equilibria may not exist, own-plan dependence may prevent the existence of pure intrapersonal equilibria for given beliefs about others or chance. The discussion of disappointment in game form G_7 (Section 3.2) provides a simple illustrative example.

Let us note here that a feature of early work on p-games and, most specifically, on guilt was the tacit assumption that a player's behavior is necessarily consistent with his plan, independently of his being rational or not. B&D, in fact, *explicitly* assume that behavior is necessarily consistent with plan and that this is transparent to the players. We can illustrate this with reference back to the discussion of guilt in Section 3.1. According to B&D, in G_6^*, when 2 observes *trust*, he takes for granted that $\hat{\pi}_1 = 10 \cdot q$ where q denotes the probability subjectively assigned by 1 to *share* (given *trust*), because 1 must have planned to trust. With this, the relevant psychological utility of 2 given (*trust,grab*) can be written as $14 - \theta_2 \cdot 10 \cdot q$. This makes a difference for the definition and implications of some solution concepts, like sequential equilibrium, as explained in detail by BC&D in their Section 7.6.

[76] See BC&D and Battigalli, Corrao & Sanna (2020).

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

[78] As in models of disappointment (Section 3.2), anger (3.3), anticipatory feelings (3.5), and belief-dependent loss aversion (5).

Deriving the rational-plan (best-reply) correspondences $\beta_i \mapsto \{\alpha_{i,i} : (15) \text{ holds}\}$ is the *first step to understand incentives and behavior* in p-games. As we explain in Section 7, in many experiments the key features of second-order beliefs are elicited and predictions rely on this step. The next step is to endogenize second-order beliefs by means of a solution concept. GP&S focused on a strong notion of sequential equilibrium, which requires that beliefs of all orders are correct and implies that players never change their mind about how others think and their intentions. B&D and BC&D generalized sequential equilibrium and also put forward other solution concepts that do not assume correct beliefs.[79]

Local utilities, incomplete information, and time Solution concepts for p-games can be defined and analyzed starting from the "local" expected 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 2), or the desire to vent one's own frustration by harming others (3.3), it is convenient to work directly with such history-dependent utility functions, without deriving them from utilities of terminal histories.

A realistic analysis of strategic thinking may have to account for uncertainty about personality traits, i.e., incomplete information. This can be achieved by parameterizing such traits with some vector θ and letting players' first-order beliefs concern the unknown part of θ as well as behavior. We argued that beliefs about personal traits—including one's own—are also essential to model some motivations such as image concerns and self-esteem (see Sections 4 and 5).[80]

Finally, as in several game-theoretic models in standard economics and behavioral economics, it may be necessary to introduce the role of time explicitly, not only to model preferences over streams of outcomes, but also to model (i) preferences over streams of emotional states (e.g., Ely *et al.* 2015), (ii) changing belief-dependent referents (e.g., Kőszegi & Rabin 2007, 2009), and (iii) decay of emotional states, such as anger, and the related actions tendencies (e.g., Gneezy & Imas 2014). To do this, one has to distinguish between **stages** and **periods**. Actions are taken within stages and information accrues to players between stages. Outcomes occur within periods, which may comprise multiple stages, and decay or discounting concern different periods. For example, we mentioned in Section 5 how Kőszegi & Rabin's concepts of Personal Equilibrium and Choice-acclimating Personal Equilibrium can be accommodated within our framework by applying the aforementioned

[79] See also Battigalli, Corrao & Sanna (2020) and Jagau & Perea (2018).

[80] The general PGT framework put forward by BC&D emphasizes incomplete information. This is also analyzed by Attanasi, Battigalli & Manzoni (2016) and Bjorndahl, Halpern & Pass (2020), who focus on the Bayesian equilibrium concept, and by Battigalli, Corrao & Sanna (2020) who focus on epistemic foundations and rationalizability.

notion of rational planning to different time frames. More generally, it is in principle important to be able to use formalism that *accurately* represents timing, sequences of moves, and information flows, because details that can be neglected in traditional game theory (e.g., the presence of dummy nodes at which players are not active) may matter for psychological considerations. Which details can be innocuously neglected, or altered should be suggested by the psychological motivations being modeled.

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

Further remarks on modeling Can and should we study the psychological phenomena we have described using traditional game theory instead of PGT? Sometimes p-games, most notably perhaps those involving image concerns as described in Section 4, can be turned into "strategically equivalent" traditional 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. As long as i believes in j's rationality, the strategic analysis of the p-game and such an associated traditional game are equivalent.[81]

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. However, the possibility of such transformations has also engendered the claim that PGT is, after all, not needed: choosing different assumptions about utility one can go back to good, old, familiar game theory, making everybody feel at home. We are critical of such attitudes. They confuse formalism with reality. The reality is described 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 reality, coming up with a false representation of reality to claim representability with an old framework can be misleading.[82]

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

[82] Furthermore, nobody has shown that all interesting forms of p-games can be turned into "equivalent" standard games. Considering claims made at seminars we have given, we suspect this is not for lack of trying. The only article we know of dealing with the topic is that of Kolpin (1992). He limits attention to the class of p-games considered by GP&S. In our view, while his exercise is pioneering and useful as an attempt at proof-of-concept, the specific assumptions he engages are too convoluted to be practical.

Differences with GP&S Our perspective and formal analysis differs from that of GP&S in several ways. Let us first address the most important for practical purposes and least important from a conceptual standpoint: unlike GP&S (and B&D), we focus on the case where *only (first- and) second-order beliefs matter for expected* (psychological) *utility calculations.* To our knowledge, this is enough to encompass the overwhelming majority of applications. Moving on to conceptually important differences, GP&S consider only *initial* beliefs about the behavior and the *initial* beliefs of *others.* In particular, in game forms with simultaneous moves (where $Z = A := A(\varnothing)$) GP&S consider utilities of the following form: $\hat{u}_i\left(a, \beta_{i,-i}^{\varnothing}\right)$, where $\beta_{i,-i}^{\varnothing} \in \Delta\left(A_{-i} \times \left(\times_{j \neq i} \Delta\left(A_{-j}\right)\right)\right)$ denotes i's initial belief about the behavior and the (first-order) beliefs of co-players. We obtain such functional forms in the special case where only initial beliefs 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 models of sequential reciprocity (Section 2), image concerns (4), deception (4.1), and self-esteem (5). Second, it rules out own-plan-dependent utility as in models with belief-based reference-dependence (Sections 3.2, 3.3, and 5) and anticipatory feelings (3.5). Third, as mentioned above, GP&S' framework restricts the toolbox of strategic analysis to (extensions of) traditional equilibrium concepts whereby players have correct beliefs about the (initial) beliefs of others, which therefore never change as play unfolds, on or off the equilibrium path. Indeed, if this were not the case (as in appropriate versions of rationalizability, see BC&D), it would be necessary to address the issue of how players update their beliefs concerning what they care about, i.e., others' beliefs.

7 Experiments

In developing theory, our favored approach is to focus on interesting assumptions about forms of belief-dependent motivation, and to explore what they imply. Our main goal is not to explain data (although we are not saying that doing so cannot make sense). With this outlook, once a theory is formulated, a natural next step is to be inquisitive as regards its empirical support. Theories formulated using PGT can be tested for empirical relevance in lab experiments. Our main goal in this section is to give a fairly exhaustive review of aspects of experimental design that are of particular relevance. We do not try to give an exhaustive account of results, although we briefly describe some prominent findings.

We cover a series of themes we deem important. For clarity, we give each a separate heading. The first themes are closely related. We list them next to each other. Later on, the themes are less closely related. We apologize that the transitions then may feel choppy.

Belief elicitation Models formulated using PGT suggest ways that particular beliefs impact preferences and play. To conduct lab tests it is often helpful to elicit those beliefs. The very first experiment specifically designed to test a PGT-based prediction was built around that insight. Dufwenberg & Gneezy (2000) considered versions of G_1 (recall: player 2 chooses $t \in \{0, ..., M\}$) as well as Trust Games where 1 could take an outside option (choose *out*) or choose *in* and let 2 choose in a subgame structured like G_1 ("Lost Wallet Games"). They 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).[83] The test for guilt checks whether for subjects in the position of player 2 there is positive correlation between t and those guess-guesses. Dufwenberg & Gneezy performed such tests and, by and large, found support for the theory.

There is a large follow-up literature testing the empirical relevance of guilt in various game forms, often eliciting beliefs. Mostly, binary Trust Games like G_6 are explored. See Cartwright (2019) for a survey.[84]

Some studies elicit beliefs in order to study other forms of motivation than guilt, again to a large degree reporting support. The pioneer to do this for reciprocity theory is Dhaene & Bouckaert (2010), who carefully elicit precisely the belief-data that the task demands (including particular conditional beliefs).[85] And a few recent studies testing aspects of Battigalli, Dufwenberg & Smith's models of frustration and anger also do it.[86]

Some scholars we met, mainly of decision-theoretic bent, seemed skeptical on grounds

[83] The description is precise as regards G_1. In the 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 eliciting the right belief, namely 2's belief conditional on 1 choosing *in*.

[84] Many scholars report continued support for the theory, though this is not universal. See also Guerra & Zizzo (2004), Charness & Dufwenberg (2006, 2010, 2011), Bacharach, Guerra & Zizzo (2007), Vanberg (2008), Miettinen & Suetens (2008) Reuben, Sapienza & Zingales (2009), Ellingsen, Johannesson, Tjøtta & Torsvik (2010), Bellemare, Sebald & Strobel (2011), Chang, Smith, Dufwenberg & Sanfey (2011), Dufwenberg *et al.* (2011), Amdur & Schmick (2013), Beck, Kerschbamer, Qiu & Sutter (2013), Bracht & Regner (2013), Kawagoe & Narita (2014), Morrell (2014), Regner & Harth (2014), Andrighetto *et al.* (2015), Khalmetski *et al.* (2015), Hauge (2016), Khalmetski (2016), Woods & Servatka (2016), Balafoutas & Fornwanger, Balafoutas & Sutter (2017), Bellemare, Sebald & Suetens (2017), Attanasi, Battigalli, Manzoni & Nagel (2019), Attanasi, Rimbaud & Villeval, Danilov *et al.* (2019), Dhami *et al.* (2019), Di Bartolomeo, Dufwenberg, Papa & Passarelli (2019), and Inderst, Khalmetski & Ockenfels (2019).

[85] See also Dufwenberg *et al.* (2011) and Attanasi, Battigalli *et al.* (2013, 2019). Results indicate that many, but not all, subjects conform with the theory.

[86] See Aina, Battigalli & Gamba (2020) and Dufwenberg, Li & Smith (2018). Also Persson (2018) performs such a test, although he does so without eliciting beliefs. With the exception of Persson not documenting support for "simple anger," much of this evidence is supportive. See also Battigalli, Dufwenberg & Smith's (Section 5) discussion of several other older experiments which were not designed to test the theory and yet may be viewed in such a light ex post.

of principle to the idea of belief-dependent motivation, and in the sequel to the idea of eliciting beliefs. This would be people who revere "revealed preference", who argue that beliefs are not real, or at least not observable. Beliefs are merely a theory feature, something that should be viewed only as part of a preference "representation". Conceivably, it is then argued that belief elicitation is pointless, as one then measures something that is not real. In our view, this position has little merit. It militates against introspection, against the rationale involved in humans' use of language, and against some of the experimental evidence we cited. Much like psychologists have by and large given up on their analogous "behaviorist" approach (favored in much of the twentieth century), many calls for revealed preference in economics seem obsolete to us. That said, however, one has to admit that there are many thorny methodological issues surrounding how to best measure subjects' beliefs.[87] Different PGT-related papers take different approaches and some (e.g., Cartwright) discuss pros & cons. See Schotter & Trevino (2014) for a (broader than just PGT) critical survey of the literature on belief elicitation in laboratory experimental economics.

Belief disclosure Charness & Dufwenberg (2006) 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 so that she believes others believe she made that choice. This would imply that a subject's choice drives his SOB, rather than the other way around (as the guilt story has it). Ellingsen, Johannesson, Tjøtta & Torsvik (2008) propose a clever alternative design, which avoids that issue but has another problem. Rather than elicit 2's SOB they elicit 1's FOB, which they *disclose* to 2 before she chooses. This induces 2's SOB without the risk of false consensus. The drawback, however, is a potential loss of control. In Ellingsen *et al.*'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 would believe 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 would affect her behavior.[88]

[87] 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).

[88] This line of criticism made Ellingsen *et al.*'s approach controversial. Yet the technique has come to be frequently relied on. See, e.g., Attanasi, Battigalli *et al.* (2013, 2019), Khalmetski *et al.*, Bellemare *et al.* (2017), Dhami *et al.*, and Danilov *et al.*

No elicitation It is not always necessary to elicit beliefs to meaningfully test PGT-based hypotheses. Sometimes patterns of behavior are idiosyncratic enough to a specific theory that clear conclusions can be drawn by observing choices only. To illustrate, recall the perceived cheating aversion theory of Section 4.1. In their experiment, using a design matching G_{10}, Fischbacher & Föllmi-Heusi found reporting frequencies fell *in between* what would obtain with honest choices (16.7% for each y) and selfish reporting (100% $y = 5$). One does not need to elicit beliefs to see that this is in line with the theoretical prediction we described.[89] A further striking insight concerns treatments that manipulate player 2's information. Recall our discussion of how (with perceived cheating aversion) such a change is predicted to undermine the partial lies prediction. Gneezy *et al.* (2018) ran such treatments, where player 2 were given information about both x and y, and report that 1's behavior indeed changed in the direction of all-or-nothing lies.

Other cases where belief-elicitation was not necessary include Charness & Dufwenberg's (2011) tests regarding "guilt-from-blame" (note especially their remark at the top of p. 1231); Dufwenberg *et al.*'s (2013) test of negative reciprocity in hold-up problems; and tests that involve a single active player and where the relevant beliefs are pinned down by chance moves—examples include tests of Kőszegi & Rabin's theory as pioneered by Ericson & Fuster (2011) and Smith (2019, but written contemporaneously) and Persson's (2018) test of Battigalli, Dufwenberg & Smith's theory.

There is also a large experimental literature which discusses and tests (and often finds support for) more informally formulated notions of reciprocity. Typically, no explicit connection to PGT is made. We will not discuss details, but see Fehr & Gächter for an early highly influential discussion.

Which theory wins? That's the wrong question! It is true that to the extent that we have so far cited experimental results, these have mostly been supportive. Someone may wonder: can really each theory, whether it concerns guilt or reciprocity or anger, be supported? And shouldn't we figure out which one is more relevant? Our answers are *yes* and a *qualified no*. We find it plausible that all (or most) of the many motivations we have discussed matter (including, e.g., all the emotions list cited at the end of Section 3.6). Why else would humans have come up with words to describe them?

As regards experiments, the style we focus on in this section is not trying to find a "best" theory for explaining data. Rather, we explore the empirical relevance of propositions that are mostly not mutually exclusive. Do we believe all of these motivations to always be relevant in all situations at the same time? No we certainly don't, and we would like to

[89] More precisely, 35% choose $y = 5$, 25% choose $y = 4$, and all other reports occur with positive frequency that declines with y. Dufwenberg & Dufwenberg explain how this pattern conforms particularly well with an equilibrium they call "sailing-to-the-ceiling.".

propose that an important research area of high (but seemingly largely untapped) potential is to figure out which situational cues trigger what kind of motivations.

Communication Charness & Dufwenberg (2006) argue that guilt can help explain why *communication*, and in particular *promises*, can foster trust & cooperation. Recall our discussion in Section 3.1. They designed an experiment to test that hypothesis, using methods similar to those of Dufwenberg & Gneezy described above. Vanberg (2008) argued that the 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, 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, only the promisor was told (not the promisee). The key idea is that promisors would suffer expectations-based guilt independently of whether or not a switch occurred, whereas a cost of breaking a promise would apply only with no switch. The commitment-based theory is *not* PGT-based. However, discussions of it typically involve comparisons with Charness & Dufwenberg's belief-based account, so it is important for PGT-scholars to know about Vanberg's work.

The studies cited in the previous paragraph report support for the story they set out to test, and lots of follow-up work has attempted to evaluate which one has stronger pull on data. It is fair to say that some studies support one while others support the other, and that a debate is ongoing. For more information, see Section 4 in Cartwright (2019) and the discussion in Di Bartolomeo *et al.* (2019).

Exogeneity & causal inference Vanberg's approach is important also for a methodological reason: Testing for belief-dependent preferences by comparing subjects who self-report different beliefs, as Charness & Dufwenberg did, has the drawback of not relying on exogenously created variation. Subjects are not randomly assigned to their (so-to-say, home-grown) 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 in 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 Charness & Dufwenberg's hypotheses.

Ederer & Stremitzer (2017) developed a design that involves exogenous variation in subjects' SOB's, and Di Bartolomeo *et al.* developed a design that features exogenous

variation in both SOB's and promises. We refer to these studies for more information, while noting that exogenous variation and causal inference has become of high importance in this literature.

Avoidance For certain PGT-related testing purposes it may be useful to employ designs that allow a subject to avoid making another subject aware of a game being played. This would presumably be useful if one were to test ideas that directly involved unawareness, like in Nielsen & Sebald's "party example" of section 5, but it can also be useful for testing whether subjects care about image as described in section 4. Consider, e.g., the design of Dana, Cain & Dawes (2006): Subject i were given a choice whether to "exit" a \$10 Dictator Game form (like G_1, with $M = 10$) and take \$9 instead, knowing that the exit option would leave receiver j nothing *and* ensure that j never knew that a dictator game form could have been played. The design provides a test whether i cares for his image. The idea is that by exiting i may enjoy a (rather) high payoff without suffering a bad image.[90]

Alternative designs can test similar hypotheses without leaving players unaware of strategic possibilities, by allowing a player to avoid revealing a choice. For example, the design of Andreoni & Bernheim (2009) examines "an extended version of the Dictator Game in which (i) chance sometimes intervenes, choosing an unfavorable outcome for the recipient, and (ii) the recipient cannot observe whether chance intervened" (p. 1609).[91] The idea is that by choosing directly the "unfavorable outcome" that chance might implement, a subject can avoid a bad image.

Method of play PGT provides a perspective to elucidate the costs and benefits of the methods of play used in experiments about sequential games and decision problems: the **direct-response method**, whereby subjects play sequentially (thus generating limited and unbalanced data about responses to different actions of early movers), and the **strategy method**, whereby subjects commit in advance (but covertly) on their contingent choices, i.e., they play the game in strategic form.[92] When and how should we expect the strategy method to distort behavior compared to what would occur in a sequential situation? This depends on the game form and the motivations that are supposed to drive behavior.

[90] An exit choice gives \$-payoff combinations $(9, 0)$ to i and j, whereby i reveals a preference for that outcome over each of the combinations $(10, 0)$, $(9, 1)$, and $(5, 5)$ which he could have obtained by not exiting. This contradicts the models of social preferences of Fehr & Schmidt, Bolton & Ockenfels, and Charness & Rabin.

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

[92] See Brandts & Charness (2011) and the references therein.

From a theoretical viewpoint, playing with the strategy method should be equivalent to playing with the direct response method when preferences are dynamically consistent. If instead, preferences are dynamically inconsistent, the ability to "tie the hands of one's later selves" by committing to a conditional response may significantly affect behavior in some situations. There are a wealth of psychological reasons for dynamic inconsistency, and PGT can help highlight when they may be relevant. To illustrate, if subjects play an Ultimatum Game, anger and inequity aversion (a non-belief-dependent motivation) both provide potential explanations for rejections. According to the theory of frustration and anger of Battigalli, Dufwenberg & Smith, anger can only arise when a responder actually observes a greedy offer; instead, the mere contemplation of such a possibility with the strategy method is unlikely to engender enough frustration and anger to make him reject.[93] By contrast, if the only reason for rejections were inequity aversion, a consequentialist model of distributional preferences,[94] then the strategy method should be equivalent to the direct response method.

For example, in the Ultimatum mini-game G_5, the commitment to *reject* the *greedy* offer makes a difference for the distribution of monetary payoffs only if the proposer indeed makes the *greedy* offer; therefore, the ex ante preference between committing to *accept* or *reject* coincides with the post-offer preference between the payoff pairs $(0, 0)$ (*reject*) and $(9, 1)$ (*accept*, with \$1 for the responder).[95] In the Trust mini-games G_6 and G_{11}, second movers cannot be negatively surprised when they have to choose. Thus, even if they are prone to anger, such emotion cannot affect behavior. Guilt aversion and image concerns, instead, are likely drivers of behavior. But, as explained in Section 6, in such models psychological utility does not depend on the agent's plan, and the resulting belief-dependent preferences are dynamically consistent. Thus, the strategy method should be equivalent to the direct response method.[96]

Other forms of data It may be useful to consider other kinds of data than choices and elicited beliefs to test PGT-based hypotheses. For example, brain imaging data (e.g., fMRI), emotion self-reports ("please rate how strongly you feel emotion X on a scale..".), electrodermal activity, or face-reader data may be useful. Chang, Smith, Dufwenberg & Sanfey (2011) pioneered the use of fMRI for PGT-related purposes, in a study taking the theory of simple guilt to the brain scanner. Chang *et al.*'s study also involved emotion

[93] See Remarks 1-2 in Battigalli, Dufwenberg & Smith.

[94] See, e.g., Fehr & Schmidt or Bolton & Ockenfels.

[95] Aina *et al.* report that the rejection rate in the Ultimatum mini-game is indeed higher with the direct response method, consistently with the hypothesis that anger is a driver of behavior.

[96] For this reason, Attanasi, Battigalli *et al.* (2013, 2016, 2019) analyze a simultaneous-move version of the Trust mini-game.

self-reports, in a way that was mindful of the possibility that pangs of guilt might be counterfactual and yet crucial (see Section 3.1 above).[97] 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) use the technology to explore anger and Battigalli, Dufwenberg & Smith cite their results when motivating their own theory.

8 Applications

We hope to inspire applied work. One may think of applications on two levels. The broader one would be to formulate within PGT a model of some belief-dependent motivation that can be used to analyze some class of game forms. We already discussed this topic, for models involving reciprocity, guilt, regret, and anger. Our focus in this section will instead be to applications of a more targeted form, exploring some particular economic setting or question. Our first and third examples of the introduction provide two specific examples, tipping and fashion choice being the economic phenomena scrutinized. Most applied work involves reciprocity, guilt, belief-dependent loss aversion, and image concerns. We listed relevant references in previous sections, but space constraints make it impossible to describe details. Instead, in this section, we attempt to provide some perspective by commenting on possible *angles* that applied research may follow.

First, one may find settings in which some form of belief-dependent motivation is plausibly operational, but which economists gave scant attention. Likely, economists may not have been used to exploring the sentiment in question, and therefore didn't look at the overall scenario. An example of such research could be Caplin & Leahy's (2004) exploration of how a caring doctor should disclose (or not) information to a patient who is about to undergo an operation and who is affected by anxiety.

Second, one may take a well-known classical model and explore whether predictions are robust to the incorporation of some form of belief-dependent motivation which is plausible in that setting. Patel & Smith (2019) can exemplify. They consider a classic setting of public goods provision and ask how the equilibrium set is affected when players are influenced by guilt.

Third, one may explore whether comparative statics that hold under classical analysis are relevant also if some form of belief-dependent motivation is at play. Dufwenberg *et al.*'s (2013) study of "the hold-up problem" can exemplify. The analysis involves a form of

[97] Chang *et al.* 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."

reciprocity theory where players only reciprocate negative kindness, and involves the game form G_{12}:

$$[G_{12}]$$

The parameter ω takes different values depending on how details of a hold-up problem are described,[98] but under classical assumption that does not matter. If utility equals (own) money, then the backward induction solution is independent of ω! However, if player 1 is inclined to take revenge, applying the theory, he will wish to choose *reject* if ω is low enough, so the hold-up problem is mitigated in these cases. The authors discuss how this may have implications for how to organize firms, a topic discussed in the academic discipline called *Strategy*.

Fourth, one may observe some pattern of behavior that looks like a puzzle from a classical viewpoint, and explore whether some form of belief-dependent motivation could plausibly be involved and resolve the puzzle. One example of such work could be the study by Conconi *et al.* (2017), which starts with the arguably puzzling observation that trade disputes are often initiated by incumbent political candidate shortly before elections (e.g., by Obama in 2012). Conconi *et al.* show that if the electorate is motivated by reciprocity then the observed pattern may be predicted. Another example is the demonstration by Herweg, Muller & Weinschenk (2010) that if agents in a contract-theoretic setting are influenced by belief-dependent loss aversion *à la* Kőszegi & Rabin, then under some circumstances optimal incentive schemes will have a simple form, involving only a fixed wage and a bonus. This addresses the puzzling observation that many real-world contracts look that way, whereas traditional contract theory predicts more complexity and nuance in regards to how payment depends on performance.

Fifth, one may demonstrate how theoretical tools that are known to have certain properties in traditional games may somehow have different implications when a belief-dependent motivation is at play. Bernheim (1994) can demonstrate what we have in mind. Besides shedding light on an image concern and conformity, his study also reveals some new insights on signaling games. Bernheim applies a particular refinement (viz., the "D1 criterion"), argues that it has proven "hostile to pooling in a variety of contexts" (p. 855), and emphasizes how special and telling it is that he nevertheless gets pooling/conformity in his

[98] The authors use the following story to derive the special case of $\omega = 2$: "An artist (player 1) has been asked by a presumptive buyer (player 2) to paint a 'beautiful portrait of 2.' 1 may disagree or agree. In the former case, 1 and 2 go separate ways. In the latter case, 1 spends $\$2,000$ worth of his/her time on the painting, and a contract says 2 should subsequently pay $\$5,000$ to 1. The value to 2 is $\$8,000$, but 2 may complain and claim (falsely) that the portrait is 'rather ugly' and attempt to renegotiate offering a new price of $\$1,000$. Given the ambiguity of what constitutes beauty, 1 cannot enforce the $\$5,000$ payment and will have to accept or reject the new offer. 1 knows that no person other than 2 would pay to acquire the painting." In another story, where the fruits of the effort becomes human capital of player 2 (which 1 cannot deny 2 access to), $\omega = 10$.

model.

Sixth, we have emphasized how in traditional games predictions do not depend on information across end-nodes, but how this ceases to be true in some p-games. One may look for interesting economic settings where this matters. An example is provided by Dufwenberg & Nordblom's (2018) study of tax evasion and guilt. It is shown that whether tax returns are private or public influences information across end-nodes, and play.

Seventh, experiments may have revealed insights regarding the impact of some belief-dependent preferences. The documentation of effects may, however, have been confined to rather special lab games, and one may explore how the ideas apply in naturally occurring economic settings. For example, consider Charness & Dufwenberg's (2006) evidence that pre-play communication, via mechanisms that depend on guilt, may foster trust and cooperation. They derive that insight in a simple binary Trust Game form. One could consider exploring whether the pattern extends also to a setting with cartelists who meet informally to discuss price-fixing.

Finally, as we wrap up this section, it is natural to wonder where the meat will be going forward. What new specific topics should one look at? It is beyond the scope of our text to exhaustively answer that question. Our goal is to inspire efforts in that direction rather than conduct new studies ourselves. Yet we cannot resist pointing in a few directions which we feel offer promising grounds for timely and original contributions.

(i) Starting from the perspective of a particular emotion, we propose that there is scope for interesting work to be done involving *anger*. In Section 3.3, we mentioned many settings including road rage and support for populist candidates. Take the last of these. Suppose that (two or many) competing non-populist candidates (or parties), somehow, make promises that, for one reason or the other, they do not live up to once elected (or at least some folks feel that way). So the elected politicians end up frustrating many voters. Now comes Donald Trump (or Matteo Salvini) and says he'll "drain the swamp," which presumably would be a policy that is annoying for the incumbent. That (populist) candidate may then get the votes of those frustrated voters just because they want to hurt the incumbent! Such a prediction would seem consistent with the model of Battigalli, Dufwenberg & Smith, although one would have to develop the application to say with more confidence.

(ii) Starting from the perspective of a particular societal topic, we believe PGT has much to offer for analyzing behavior and economic outcomes during *pandemics*. As a matter of fact, we are not the first to argue this point. It is the core idea of a recent paper by Peter Huang (2020). He "advocates building rigorous, theoretical models to analyze how leaders and leadership communications in pandemics can reduce anger, anxiety, and frustration, prevent panic, inhibit complacency, and foster compliance with NPIs" (i.e., non-pharmaceutical interventions, like social distancing and self-quarantining), and goes

on to argue that PGT provides relevant tools. We couldn't agree more, and strongly recommend his text for inspiration.

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

9 Concluding remarks

Decisions are driven by a plethora of motivations. 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.[99]

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 motivations that shape behavior in important ways are belief-dependent. This includes reciprocity, emotions, image concerns, belief-dependent loss aversion, 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 show why & how. We showcased new phenomena allowed by belief-dependent motivation and related technical issues arising in PGT: Inactive players and their information may affect the incentives of active players and outcomes due to, e.g., image concerns (see the Introduction and Section 4). Multiple non-equivalent equilibria arise where traditional game theory implies a unique equilibrium (Sections 2 and 3.2). Preferences may be dynamically inconsistent due to own-plan dependence (Sections 3.2, 5 and 6), and pure rational plans may not exist (Section 3.2). We do not see these

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

features of PGT as daunting complications, but rather as modeling opportunities. Relatedly, we argue that the often raised criticism that PGT makes almost everything possible is misconceived. Like traditional game theory, also PGT is a formal framework/language that allows to build models of interactive decision making. The art of modeling consists in abstracting from some aspects and focusing on others based on our judgement about their relevance in a given context. Specific models may be relatively simple even if they do not have standard features, and they may yield testable predictions (Section 7).

Working with PGT is exciting, and interest is growing. For example, the *Journal of Economic Behavior and Organization* recently published a special issue on PGT, comprising 15 articles (cited above); see Dufwenberg & Patel (2019) for the guest editors' introduction. We derive utility from our *hope* (=item #12 in Elster's list from Section 3.6) to inspire others to follow suit.

References

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

[2] Aina, Chiara, Pierpaolo Battigalli, and Astrid Gamba. 2020. "Frustration and Anger in the Ultimatum Game: An Experiment". *Games and Economic Behavior* 122: 150-167.

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

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

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

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

[7] Andrighetto, Giulia, Daniela Grieco, and Luca Tummolini. 2015. "Perceived Legitimacy of Normative Expectations Motivates Compliance with Social Norms when Nobody is Watching". *Frontiers in Psychology* 6, 1413.

[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. "Belief-Dependent Preferences and Reputation: Experimental Analysis of a Repeated Trust Game". *Journal of Economic Behavior and Organization* 167: 341-360.

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

[11] Attanasi, G., and R. Nagel. 2008. "A Survey of Psychological Games: Theoretical Findings and Experimental Evidence." In A. Innocenti and P. Sbriglia (Eds.) *Games, Rationality and Behavior. Essays on Behavioral Game Theory and Experiments*, Palgrave McMillan, 204-232.

[12] Attanasi, Giuseppe, Claire Rimbaud, and Marie-Claire Villeval. 2019. "Embezzlement and Guilt Aversion". *Journal of Economic Behavior and Organization* 167: 409-429.

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

[14] Azar, Ofer H. 2019. "The Influence of Psychological Game Theory". *Journal of Economic Behavior and Organization* 167: 445-453.

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

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

[17] Balafoutas, Loukas, and Helena Fornwagner. 2017. "The Limits of Guilt". *Journal of the Economic Science Association* 3: 137-148.

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

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

[20] Battigalli Pierpaolo, Roberto Corrao, and Martin Dufwenberg. 2019. "Incorporating Belief-Dependent Motivation in Games". *Journal of Economic Behavior and Organization* 167: 185-218.

[21] Battigalli, Pierpaolo, Roberto Corrao, and Federico Sanna. 2020. "Epistemic Game Theory without Types Structures: An Application to Psychological Games". *Games and Economic Behavior* 120: 28-57.

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

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

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

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

[26] Baumeister, Roy, Arlene Stillwell, and Todd Heatherton. 1994. "Guilt: An Interpersonal Approach". *Psychological Bulletin* 115: 243-267.

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

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

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

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

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

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

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

[34] Bénabou, Roland, and Jean Tirole. 2011. "Identity, Morals, and Taboos: Beliefs as Assets". *Quarterly Journal of Economics* 126: 805-855.

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

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

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

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

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

[40] Bicchieri, Cristina. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

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

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

[43] Bjorndahl, Adam, Joseph Halpern, and Rafael Pass. 2020. "Bayesian Games with Intentions". *Games and Economic Behavior* 123: 54–67.

[44] Blume, Andreas, Ernest K. Lai, and Wooyoung Lim. 2019. "Eliciting Private Information with Noise: The Case of Randomized Response". *Games and Economic Behavior* 113: 356-380.

[45] Bolton, Gary, and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition". *American Economic Review* 90: 166-93.

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

[47] Brandts Jordi, and Gary Charness. 2011. "The Strategy versus the Direct-Response Method: A First Survey of Experimental Comparisons". *Experimental Economics* 14: 375-398.

[48] Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson. 2007. "Is Generosity Involuntary?". *Economics Letters* 94: 32-37.

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

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

[51] 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.

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

[53] Caria, Stefano, Marcel Fafchamps. 2019. "Expectations, Network Centrality, and Public Good Contributions: Experimental Evidence from India". *Journal of Economic Behavior and Organization* 167: 391-408.

[54] Cartwright, Edward. 2019. "A Survey of Belief-Based Guilt Aversion in Trust and Dictator Games". *Journal of Economic Behavior and Organization* 167: 430-444.

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

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

[57] Chang, Luke, and Alec Smith. 2015. "Social Emotions and Psychological Games". *Current Opinion in Behavioral Sciences* 5: 133-140.

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

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

[60] Charness, Gary, and Martin Dufwenberg. 2010. "Bare Promises: An Experiment". *Economics Letter* 107: 281-283.

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

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

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

[64] Connolly, Terry, and David Butler. 2006. "Regret in Economic and Psychological Theories of Choice". *Journal of Behavioral Decision Making* 19: 139-154.

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

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

[67] d'Adda, Giovanna, Martin Dufwenberg, Francesco Passarelli, Guido Tabellini. 2020. "Social Norms with Private Values: Theory & Experiments". *Games and Economic Behavior* 124: 288-304.

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

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

[70] Danilov, Anastasia, Kyril Khalmetski, and Dirk Sliwka. 2019. "Norms and Guilt". CESifo w.p. 6999. CESifo Group Munich.

[71] Darwin, Francis. 1898, *The Life and Letters of Charles Darwin*. Edited by his son Francis Darwin. New York: D. Appleton and Co..

[72] Dhami, Sanjit, Mengxing Wei, and Ali al-Nowaihi. 2019. "Public Goods Games and Psychological Utility: Theory and Evidence". *Journal of Economic Behavior and Organization* 167: 361-390.

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

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

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

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

[77] Dufwenberg, Martin. 2008. "Psychological Games". In *The New Palgrave Dictionary of Economics* edited by S.N. Durlauf and L.E. Blume. Volume 6: 714-718. Palgrave Macmillan.

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

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

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

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

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

[83] Dufwenberg, Martin and Georg Kirchsteiger. 2019. "Modelling Kindness". *Journal of Economic Behavior and Organization* 167: 228-234.

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

[85] Dufwenberg, Martin, and Senran Lin. 2019. "Regret Games". Unpublished.

[86] Dufwenberg, Martin, and Michael Lundholm. 2001. "Social Norms and Moral Hazard". *Economic Journal* 111: 506-525.

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

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

[89] Dufwenberg, Martin, and Amrish Patel. 2019. "Introduction to Special Issue on Psychological Game Theory," *Journal of Economic Behavior and Organization* 67, 181-84.

[90] Dufwenberg, Martin, and David Rietzke. 2016. "Banking on Reciprocity: Deposit Insurance and Insolvency". Unpublished.

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

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

[93] Eil, David, and Justin Rao. 2011. "Asymmetric Processing of Objective Information about Yourself". *American Economic Journal: Microeconomics* 3: 114-138.

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

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

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

[97] Elster, Jon. 1996. "Rationality and the Emotions". *Economic Journal* 106: 1386-1397.

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

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

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

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

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

[103] 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.

[104] Fagin, Ronald, and Joseph Halpern. 1988. "Belief, Awareness and Limited Reasoning". *Artificial Intelligence* 34: 39-76.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[120] Gouldner, Alvin. 1960. "The Norm of Reciprocity: A Preliminary Statement". *American Sociological Review* 25: 161-178.

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

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

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

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

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

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

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

[128] Halpern, Joseph, and Rafael Pass. 2012. "Iterated Regret Minimization: A New Solution Concept". *Games and Economic Behavior* 74: 194-207.

[129] Hauge, Karen 2016. "Generosity and Guilt: The Role of Beliefs and Moral Standards of Others". *Journal of Economic Psychology* 54, 35-43.

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

[131] Herweg, Fabian, Muller, Daniel, and Weinschenk, Philipp. 2010. "Binary Payment Schemes: Moral Hazard and Loss aversion". *American Economic Review* 100, 2451-2477.

[132] Huang, Peter. 2020. "Pandemic Emotions, Public Health, Financial Economics, and Law, and Leadership". University of Colorado Law Legal Studies Research Paper No. 20-14.

[133] Huang, Peter, Ho-Mou. 1994. "More Order without More Law: A Theory of Social Norms and Organizational Cultures". *Journal of Law, Economics, and Organization* 12: 390-406.

[134] Hupkau, Claudia, and François Maniquet. 2018. "Identity, Non-Take-Up and Welfare Conditionality". *Journal of Economic Behavior and Organization* 147: 13-27.

[135] Inderst, Roman, Kyril Khalmetski, and Axel Ockenfels. 2019. "Sharing Guilt: How Better Access to Information May Backfire". *Management Science.*65: 2947-3448.

[136] Isoni, Andrea, and Robert Sugden. 2019. "Reciprocity and the Paradox of Trust in Psychological Game Theory". *Journal of Economic Behavior and Organization* 167: 219-227.

[137] Jagau, Stephen, and Andrés Perea. 2018. "Common Belief in Rationality in Psychological Games". Epicenter w.p. 10.

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

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

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

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

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

[143] Keltner, Dacher, Jennifer Lerner. 2010. "Emotions." In *Handbook of Social Psychology*, Vol. 1, edited by Susan Fiske, Daniel Gilbert and Gardner Lindzey: pp. 317-52. New York: Wiley.

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

[145] Khalmetski, Kiryl. 2019. "The Hidden Value of Lying: Evasion of Guilt in Expert Advice". *Journal of Economic Behavior and Organization* 167: 296-310.

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

[147] Khalmetski, Kiryl and Dirk Sliwka. 2019. "Disguising Lies - Image Concerns and Partial Lying in Cheating Games". *American Economic Journal: Microeconomics* 11: 79-110.

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

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

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

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

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

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

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

[155] Kozlovskaya, Maria, and Antonio Nicolò. 2019. "Public Good Provision Mechanisms and Reciprocity". *Journal of Economic Behavior and Organization* 167: 235-244.

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

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

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

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

[160] Lerner, Jennifer, and Dacher Keltner. 2000. "Beyond Valence: Toward a Model of Emotion-specific Influences on Judgement and Choice". *Cognition and Emotion* 14: 473-93.

[161] Lerner, Jennifer, and Dacher Keltner. 2001. "Fehr, Anger, and Risk". *Journal of Personality and Social Psychology* 81: 146-159.

[162] Lerner, Jennifer, Ye Li, Piercarlo Valdesolo, and Karim Kassam. 2015. "Emotion and Decision Making". *Annual Review of Psychology* 66: 799-823.

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

[164] Livio, Luca, and Alessandro De Chiara. 2019. "Friends or Foes? Optimal Incentives for Reciprocal Agents". *Journal of Economic Behavior and Organization* 167: 245-278.

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

[166] Loewenstein, George, and Andras Molnar. 2018. "The Renaissance of Belief-based Utility in Economics". *Nature Human Behavior* 2: 166-167.

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

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

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

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

[171] Marcus-Newhall, A., Pedersen, W.C., Carlson, M., Miller, N. (2000). "Displaced Aggression is Alive and Well: A Meta-analytic Review". *Journal of Personality and Social Psychology* 78, 670–689.

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

[173] Miettinen, Topi, and Sigrid Suetens. 2008. "Communication and Guilt in a Prisoner's Dilemma". *Journal of Conflic Resolution* 52: 945-960.

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

[175] Molnar, Andras, Shereen Chaudhry & George Loewenstein. 2020. "It's Not About the Money. It's About Sending a Message! Unpacking the Components of Revenge". Unpublished.

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

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

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

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

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

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

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

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

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

[185] Patel, Amrish, and Alec Smith. 2019. "Guilt and Participation". *Journal of Economic Behavior and Organization* 167: 279-295.

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

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

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

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

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

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

[192] Regner, Tobias, and N.S. Harth. 2014. "Testing Belief-Dependent Models". W.p. Max Planck Institute of Economics.

[193] 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.

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

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

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

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

[198] Sebald, Alexander, and Nick Vikander. 2019. "Optimal Firm Behavior with Consumer Social Image Concerns and Asymmetric Information". *Journal of Economic Behavior and Organization* 167: 311-330.

[199] Sebald, Alexander, and Markus Walzl. 2015. "Optimal Contracts Based on Subjective Evaluations and Reciprocity". *Journal of Economic Psychology* 47: 62-76.

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

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

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

[203] Skinner, B.F. 1948. *Walden Two*. Englewood Cliffs, NJ: Prentice Hall.

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

[205] Sohn, Jin and Wenhao Wu. 2020. "Reciprocity with Uncertainty about Others". Unpublished.

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

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

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

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

[210] Vanberg, Christoph. 2008. "Why Do People Keep Their Promises? An Experimental Test of Two Explanations". *Econometrica* 76, 1467-1480.

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

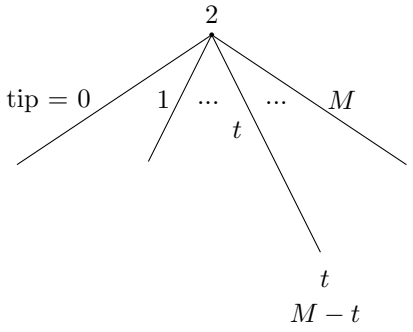
[212] van Leeuwen, Boris, Charles Noussair, Theo Offerman, Sigrid Suetens, Matthijs van Veelen, and Jeroen van de Ven. 2018. "Predictably Angry – Facial Cues Provide a Credible Signal of Destructive Behavior". *Management Science* 64: 3352-3364.

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

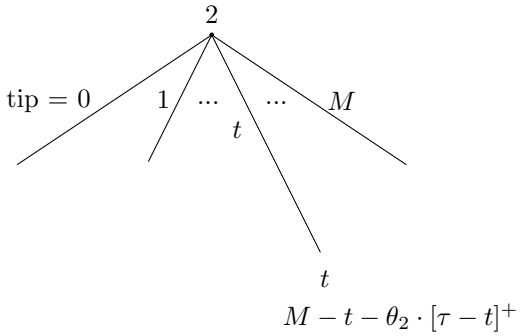
[214] Zeelenberg, Marcel. 1999. "Anticipated Regret, Expected Feedback and Behavioral Decision Making". *Journal of Behavioral Decision Making* 12: 93-106.

[215] Zeelenberg, Marcel & Rik Pieters. 2007. "A Theory of Regret Regulation 1.0". *Journal of Consumer Psychology* 17: 3-18.

G_1



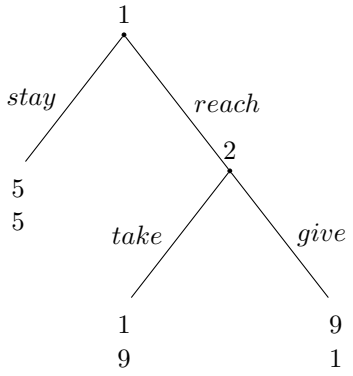
G_1^*



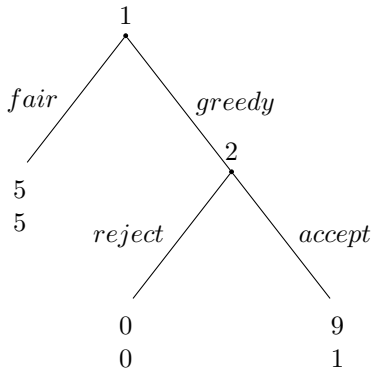
G_2

	opera	*boxing*
opera	1/2 1	0 0
boxing	0 0	1 1/2

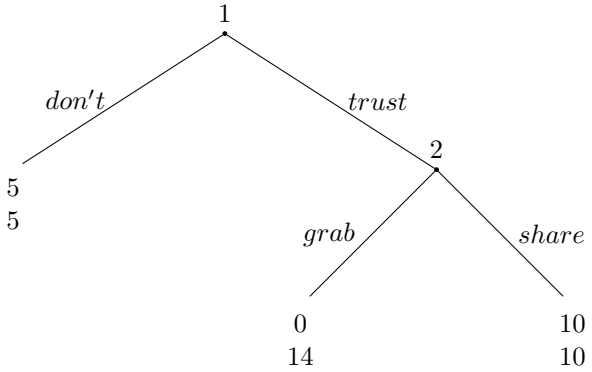
G_4


1
stay
reach
2
5
5
take
give
1
9
9
1

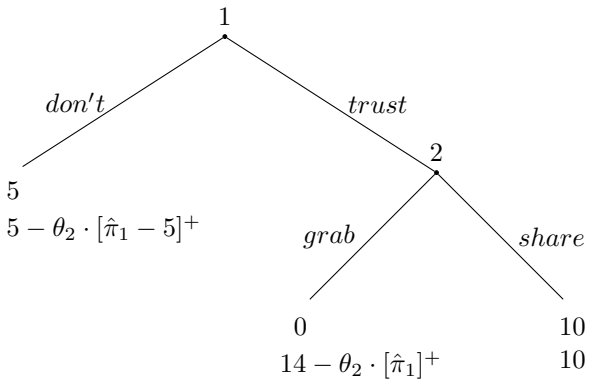
G_5


1
fair
greedy
2
5
5
reject
accept
0
0
9
1

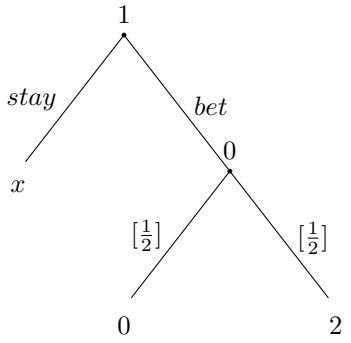
G_6


1
don't
trust
2
5
5
grab
share
0
14
10
10

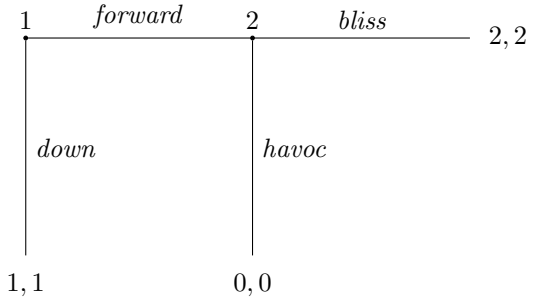
G_6^*


1
don't
trust
2
5
$5 - \theta_2 \cdot [\hat{\pi}_1 - 5]^+$
grab
share
0
$14 - \theta_2 \cdot [\hat{\pi}_1]^+$
10
10

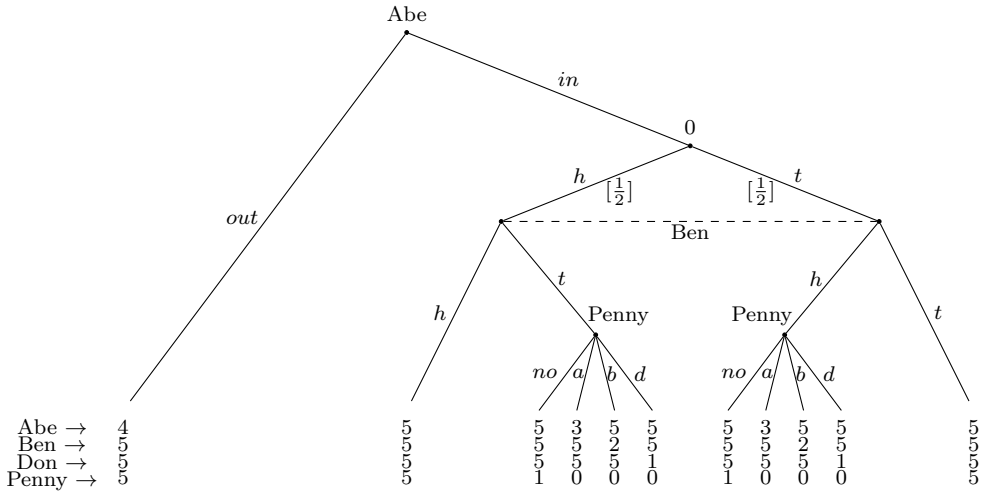
G_7


1
stay
bet
0
x
$[\frac{1}{2}]$
$[\frac{1}{2}]$
0
2

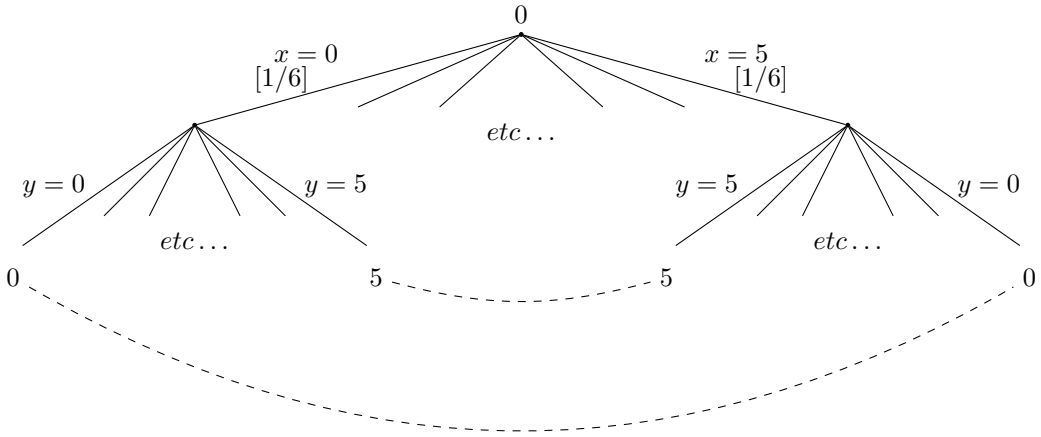
G_8


1
forward
2
bliss
2, 2
down
havoc
1, 1
0, 0

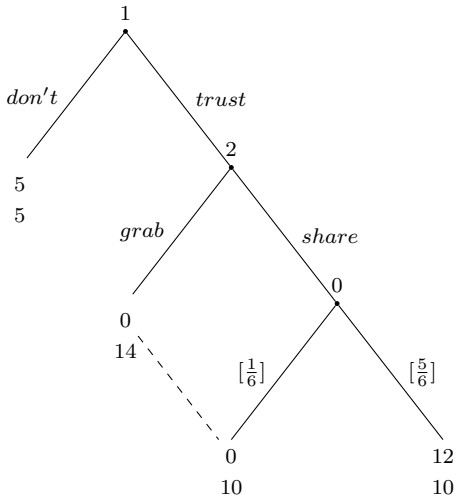
G_9


Abe
in
out
0
h
[1/2]
t
[1/2]
Ben
h
t
h
t
Penny
Penny
no
a
b
d
no
a
b
d
Abe → 4
Ben → 5
Don → 5
Penny → 5
5
5
5
5
5 3 5 5
5 5 2 5
5 5 5 1
1 0 0 0
5 3 5 5
5 5 2 5
5 5 5 1
1 0 0 0
5
5
5
5

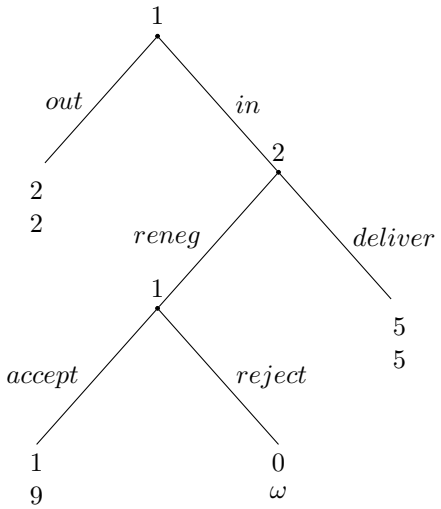
G_{10}


0
x = 0
[1/6]
x = 5
[1/6]
etc...
y = 0
y = 5
y = 5
y = 0
etc...
etc...
0
5
5
0

G_{11}


1
don't
trust
2
5
5
grab
share
0
0
14
[1/6]
[5/6]
0
10
12
10

G_{12}


1
out
in
2
2
2
reneg
deliver
1
5
5
accept
reject
1
9
0
ω