99. Psychological game theory

Pierpaolo Battigalli and Martin Dufwenberg

Psychological Game Theory (PGT) is a formal framework that generalises traditional game theory by allowing utilities to depend not only on choices but also on players' first- or higher-order beliefs about how a game is played. The framework's name is apt as it allows for analysing many forms of psychology. Namely, PGT is useful for modelling a rich host of human motivations, including emotions like disappointment, regret, guilt, frustration, anger, and fear; reciprocity, that is, the inclination to respond to kindness with kindness and to be unkind to whoever is unkind; and image concerns, as when someone desires for others, or oneself, to view oneself as, for example, brave, talented, generous, or honest.

PGT was first conceived and explored by Geanakoplos, Pearce & Stacchetti (1989) (but also compare with Gilboa & Schmeidler 1988, who presented some related ideas) and further developed by Battigalli & Dufwenberg (B&D) (2009) and Battigalli, Corrao & Dufwenberg (2019). A growing subsequent related literature develops many applications, either by exploring how to model interesting forms of psychological motivations in a class of game forms or by taking some such motivation for granted and focusing on its implications in particular economic contexts of interest. Another related literature explores the empirical relevance of the applied work using laboratory experiments. Rather than summarising all this work, we refer to the extensive recent survey article by B&D (2022). It gives a comprehensive account of the basic nature of PGT as well as the applied and experimental work.

B&D (2022) describe how to model a long list of motivations, organised as indicated above (emotions, reciprocity, image concern, and other forms of motivation). Here we provide a short alternative but complementary account that stresses the nature and timing of various belief-dependent motivations as modelled using PGT. Rather than start with a particular form of motivation and derive the nature of the relevant belief-dependence involved, we highlight a form of belief-dependence and then look for a motivation that can exemplify it. Proceeding this way, we highlight two important dimensions along which belief-dependent preferences may be classified. All of the combinations that we describe may be empirically relevant as they are reflected in common motivations.

Consider the 2×3 = 6 cases described in Table 99.1, distinguished by whether a belief dependency in player *i*'s utility refers to *i*'s own belief or to the belief of another player, as well as whether the belief in question is initial (formed at the root of a game), current (formed contemporaneously with an interim choice), or terminal (formed at the time a game ends).

Cell #1 can be illustrated by the emotion of disappointment. Player *i* is disappointed at terminal node *z* if *i*'s monetary (say) payoff at *z* is lower than the payoff m_i^e that *i* initially expected to get in the game. The size of m_i^e depends on *i*'s initial beliefs. Player *i*'s disappointment is modelled as an emotional cost that *i* incurs at *z* and which depends on m_i^e. See B&D (2022, Section 3.2) for more details and relevant related references.

Cell #2 can be illustrated by the emotion of guilt. Player *i* is affected by guilt at terminal node *z* if another player *j*'s payoff at *z* is lower than the payoff m_j^e that *j* initially expected to get. Guilt is modelled as an emotional cost that *i* incurs at *z* and which depends on m_j^e. Player *i* has no direct access to *j*'s beliefs, but *i* forms beliefs about *j*'s belief when *i* calculates expected utility. See B&D (2022, Section 3.1) for more details and relevant related references (including to B&D's 2007 general model of 'simple guilt,' which is what we have in mind here).

Cells #3 and #4 can be illustrated by reciprocity. At information set *h*, player *i*'s utility depends on *i*'s own kindness, which is shaped by *i*'s current beliefs at *h* (cell #3), as well as by *j*'s kindness, which depends on *j*'s current beliefs at *h* (cell #4). See B&D (2022, Section 2) for more details and relevant related references (including the model of Dufwenberg & Kirchsteiger (2004), which is what we have in mind here, and the related models of Rabin (1993) and Falk & Fischbacher (2006)).

Cell #5 can be illustrated by regret. At terminal node *z*, player *i*'s regret depends on *i*'s beliefs at *z* about what payoff maximum m_i^{max} that *i* believes they would have received had *i* chosen differently than they actually did earlier in the game. Regret is modelled as an emotional cost that *i* incurs at *z* and which depends on m_i^{max}. See B&D (2022, Section

Table 99.1 Classification of belief-dependent preferences in psychological game theory

-	Initial	Current	Terminal
Own	(cell#1)	(cell #3)	(cell #5)
Another's	(cell #2)	(cell #4)	(cell #6)

Note: Illustrating the classification of belief-dependent preferences in a 2×3 matrix: Player's own beliefs [initial (cell#1), current (cell#3), terminal (cell#5)] and another player's beliefs [initial (cell#2), current (cell#4), terminal (cell#6)].

3.4) for more details and relevant related references.

Cell #6 can be illustrated by models of social image. At terminal node z, player i cares about j's beliefs at z about qualities regarding i's nature (such as the shape of i's utility function, in a context where this is not known and yet inferred by j) or qualities of i's choices (such as whether i has misreported information during a game where that would be a possibility). This concern for a desirable social image is modelled as an argument in i's utility at z that depends on j's beliefs at z. See B&D (2022, Section 4) for more details and relevant related references.

When the belief-dependence of utility involves a terminal belief (such as regret or social image), information structure across terminal nodes may matter to predictions. This illustrates one of the many ways that PGT-based models sometimes have drastically different properties than traditional models. For discussions of other differences, see B&D (2022).

In all the models we mentioned so far, the belief-dependence of utility involves a first-order belief (about choices) rather than a higher-order belief (about beliefs). B&D (2022) argue that many interesting motivations have that property, but exceptions exist, see B&D's (2007) model of 'guilt-from-blame' (not the same as the 'simple guilt' model mentioned above) and Battigalli, Dufwenberg & Smith's (2019) model of 'anger from blaming intentions'.

References

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

Battigalli, Pierpaolo, and Martin Dufwenberg. 2007. "Guilt in Games." *American Economic Review* 97: 170–76.

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

Battigalli, Pierpaolo, and Martin Dufwenberg. 2022. "Belief-Dependent Motivations and Psychological Game Theory." *Journal of Economic Literature* 60: 833–82.

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

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior* 47: 268–98.

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

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological Games and Sequential Rationality." *Games and Economic Behavior* 1: 60–79.

Gilboa, Itzhak, and David Schmeidler. 1988. "Information Dependent Games: Can Common Sense Be Common Knowledge?" *Economics Letters* 27: 215–21.

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