

Psychological Games*

Martin Dufwenberg

June 2006

Abstract: Traditional game-theoretic models assume that utilities depend only on actions. This is not sufficient for describing the motivations and choices of decision makers who care about reciprocity, emotions, or social rewards. Psychological games allow utilities to depend directly on beliefs (about beliefs) besides which actions are chosen, and they can capture a wider range of motivations. This Palgrave entry contains several examples and it is indicated where research on psychological games is headed.

Suggested JEL codes: C7, D8

Traditional game-theoretic models presume that utilities depend on actions. While this framework is quite general (it can, for example, accommodate profit-maximization, altruism, inequity aversion, and Rawlsian maximin preferences) it is not rich enough to adequately describe several psychological or social aspects of motivation which depend directly on beliefs (about beliefs) besides which actions are chosen. The following example illustrates:

Karen feels guilty if she lets others down. When paying her landscaper (Jim), this influences her tipping. The more she believes Jim believes he will receive as a tip, the more she gives. More precisely, she gives just as much as she believes Jim believes he will get, in order to avoid the feelings of guilt that will plague her if she gives less.

Beyond depicting something arguably realistic, the example illustrates in the simplest possible way how one may have to transcend traditional game theory to model a belief-dependent motivation: Consider a standard game form where Karen chooses a tip t such that $0 \leq t \leq w$, where w is the number of dollars in her wallet, and where the landscaper has no choice (his strategy set is modeled as a singleton $\{x\}$). Karen's choice of tip thus pins down a strategy profile (t, x) . In traditional game theory, payoffs are defined on strategy profiles (or on endnodes

* Prepared for: S. N. Durlauf and L. E. Blume, *The New Palgrave Dictionary of Economics*, forthcoming, Palgrave Macmillan, reproduced with permission of Palgrave Macmillan. This article is taken from the author's original manuscript and has not been reviewed or edited. The definitive published version of this extract may be found in the complete *New Palgrave Dictionary of Economics* in print and online, forthcoming

induced by strategy profiles), so Karen's best choice (or choices) would be independent of her belief about Jim's belief about her choice of tip. This runs counter to the example.

Gilboa & Schmeidler (1988) and Geanakoplos, Pearce & Stacchetti (1989) present several examples that illustrate the inadequacy of traditional methods to represent preferences that reflect various forms of belief-dependent motivation. Geanakoplos *et al* develop a new analytical framework, in which the centerpiece is the notion of a *psychological game*, which may be seen as a generalization of a traditional game and which can model some of the desired effects. A psychological game differs from a traditional game in that utilities are defined on beliefs (about actions and beliefs), as well as on which actions are chosen. (The term "game with belief-dependent motivation" would be more descriptive than the term "psychological game," but I stick with the latter which has become established.)

The most well-known example of a psychological-games based application is Rabin's (1993) highly influential model of reciprocity, according to which players wish to act kindly (unkindly) in response to kind (unkind) actions. The key notion of kindness depends on beliefs in a way such that reciprocal motivation can only be described using psychological games. To see why, suppose that I jump out in front of your car blocking your way, so that you can't cross a bridge and you therefore arrive late to an important meeting. Am I kind? Clearly one cannot say without knowing what my beliefs are. If I believe the bridge is as sturdy as bridges usually are and I am just goofing around, then I am unkind. However, if I believe the bridge is about to collapse then I am kind. Arguably, I would be kind even if I mistook a sturdy bridge for a dangerous one. So, should you be kind or unkind in return? The answer depends on your beliefs about my kindness, and hence on your beliefs about my beliefs. It takes a psychological game to model that. (The example given here is similar in spirit to another example given on p 23 of Rabin (1998). Rabin's model is normal-form based. See Dufwenberg & Kirchsteiger (2004) for an extension to extensive game forms. See Fehr & Gächter (2000) and Sobel (2005) for general discussions of why reciprocity has important economic consequences.)

Reciprocity is but one form of motivation that can be modeled by means of psychological games. Many emotions are good candidates. In his article "Emotions and Economic Theory", Elster (1998) argues that a variety of emotions have important economic consequences and he laments the attention economists have paid to this. He argues that a key characteristic of emotions is that "they are triggered by beliefs" (p. 49). He 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. He asks (p. 48): “[H]ow can emotions help us explain behavior for which good explanations seem to be lacking?” Psychological games may be useful for providing answers.

Little work has been done though. One exception is Caplin & Leahy’s (2004) health care model in which a physician is concerned with a patient’s belief-dependent anxiety (cf. also Caplin & Leahy 2001). Another exception is the emotion of guilt for which a string of results, both theoretical and experimental, have been established for the specific context of trust games (see Huang & Wu (1994), Dufwenberg (1995, 2002), Dufwenberg & Gneezy (2000), Bacharach, Guerra & Zizzo (2004), Charness & Dufwenberg (2005)). I shall elaborate in some detail on these latter findings (borrowing eclectically from the cited works), since they may be suggestive of the importance of psychological games more generally in a variety of ways.

Consider the game in Figure 1, where payoffs reflect money income (first for player A, then for player B) but not the players’ preferences which may depend also on guilt as will be indicated.

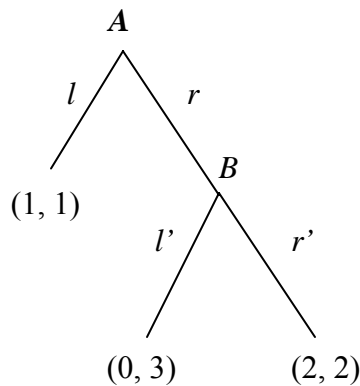


Figure 1

Assume that the more strongly player B believes that player A believes that B will make choice r' , the more guilt B would feel making choice l' and the more likely B is to make choice r' . Specifically, the players’ utilities at the various end nodes in the game form of Figure 1 coincide with the monetary payoffs, *except* following the choice sequence (r, l') where B’s utility is $3 \cdot (1 - \beta)$ rather than 3, and where β is a measure of B’s belief (with range from 0 to 1) about A’s belief that B will choose r' . (More specifically, B has a probability measure describing her beliefs about which probability A assigns to the choice r' conditional on A choosing r ; β is the

mean of that measure.) Say that B is *guilt averse*. This is all modeled in the psychological game in Figure 2:

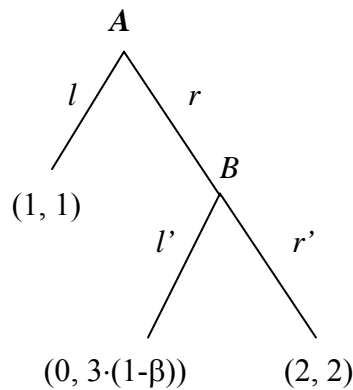


Figure 2

I wish to make several points: First, the guilt aversion modeled in Figure 2 is similar to that involved in the above example featuring Karen & Jim. In fact, the idea that people feel guilty in proportion to the degree to which they do not live up to another's expectations could be extended to any game.

Second, one can test for guilt aversion experimentally, but this requires one to measure B 's belief β . This can be done by inviting subjects to make guesses about one another's choices and guesses, rewarding accuracy in the guesswork. Such experimental tests have indicated that the prediction of guilt aversion is empirically supported in trust games. (The involved form of belief elicitation could conceivably be usefully complemented by two other forms of measurement: emotional self-reports and neurological methods such as fMRI.)

Third, guilt aversion may provide the seeds of a theory why communication can help foster trust and cooperation. To illustrate with reference to Figure 2, suppose that, before play, A and B meet and talk. Player B looks player A in the eye and *promises* to choose r' . If A believes this, and if B believes that A believes this, then guilt aversion would make B live up to her promise. A promise by B can thus feed a self-fulfilling circle of beliefs about beliefs that r' will be chosen. In combination with guilt aversion, words may be tools that create commitment power, which may in turn foster trust and cooperation.

Fourth, even without communication between A and B , one may argue that if B is guilt averse (as described above) then trust and cooperation will ensue. If A is rational and maximizes his expected monetary income (recall: we have assumed that A is selfish in this way) then by

choosing r he *signals* a certain strength of belief in B choosing r' ; if A did not assign a probability of at least $\frac{1}{2}$ to B choosing r' then he would rather choose l . If B figures this out, it puts a lower bound of $\frac{1}{2}$ on β . So, B is forced to hold a belief such that she would feel so guilty if she chose l' that she prefers r' ; in numbers: with $\beta \geq \frac{1}{2}$ we get $3 \cdot (1 - \beta) < 2$. If A figures this out, he should of course choose r . The illustrated phenomenon has been labeled *psychological forward induction*.

To sum things up: the idea of guilt aversion is intuitively plausible, experimentally testable, empirically supported, relevant for explaining why communication matters to economic behavior, and suggestive of intriguing signaling issues that may shape emotions and behavior. These insights concern a very special emotion and a very special psychological game, but seem profound given that limited scope. One may reasonably suspect that exciting conclusions are in store also for other emotions and other strategic settings, and that these conclusions may in part concern communication or belief signaling.

The discussion up till now may have been misleading in its rather heavy emphasis on reciprocity and emotions. Psychological game theory may be relevant also for describing certain social rewards (norms, respect, and status), where decision makers somehow care about the opinions or views of others. Bernheim (1994) and Dufwenberg & Lundholm (2001) present models that bear this out. These authors do not make explicit mention of psychological games, but if one takes a close look at the mathematical details one can discover connections.

One might hope that Geanakoplos *et al*'s framework is appropriate for tackling all the interesting problems that psychological games may be relevant for. However, this is not the case. A careful scrutiny reveals that their approach is too restrictive to handle many plausible forms of belief-dependent motivation (as they acknowledge themselves; see pp. 70, 78). There are several reasons, including the following:

- (1) Geanakoplos *et al* only allow initial beliefs to enter the domain of a player's utility, while many forms of belief-dependent motivation require *updated* beliefs to play a role.
- (2) Geanakoplos *et al* only allow a player's own beliefs to enter the domain of his utility, while there are conceptual and technical reasons to let *others'* beliefs matter.
- (3) Geanakoplos *et al* follow the traditional extensive-games approach of letting strategies influence utilities only insofar as they influence which end node is reached,

but many forms of belief-dependent motivation become compelling in conjunction with preferences that depend on strategies in ways not captured by end nodes.

(4) Geanakoplos *et al* restrict attention to equilibrium analysis, but in many strategic situations there is little compelling reason to expect players to coordinate on an equilibrium and one may wish to explore alternative assumptions.

(1) is manifest, for example, in the above psychological forward induction argument which hinges crucially on B's motivation depending on an updated belief. (2) is relevant, for example, for modeling social rewards (cf. the above comments on Bernheim's and Dufwenberg & Lundholm's models). As regards (3), one can show that the issue comes up if one wants to model, for example, regret, disappointment, or guilt. (4) echoes considerations relevant also for traditional games; equilibrium play is not a self-evident proposition in many contexts, for example if one assumes (only) that there is common belief in rationality, or in learning scenarios.

The list (1)-(4) is adapted from Battigalli & Dufwenberg (2005), who elaborate in more detail on each issue and take first steps towards developing psychological game theory in the indicated directions. Their approach draws crucially on Battigalli & Siniscalchi's (1999) work on how to represent hierarchies of conditional beliefs.

The decision-theoretic foundations of psychological game theory are not well understood. Classical decision theory (say von Neumann/Morgenstern) does not apply straightforwardly. To see this, take the emotion of disappointment as an example. It is plausible that disappointment is a belief-dependent emotion. To exemplify, I have 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 extremely lucky drawing to an inside straight, and then he hit his card. Another example could be based on the lotteries used in the so-called Allais' paradox. In both cases the level of disappointment, which if anticipated might affect choice behavior, may plausibly depend on the strength of prior belief that a decision maker will win a lot of money. It follows that, unless consequences are described so as to include a specification of disappointment, the so-called "independence axiom" will not make much sense for decision makers who are prone to disappointment.

Decision-theorists have not given related matters zero attention (though it seems scant). Machina (1981, pp 172-3; 1989 p 1662) presents examples in spirit related to the one million dollar example above. Bell (1985), Loomes & Sugden (1986), Karni (1992), and Karni & Schlee

(1995) go on to develop models in which utility may depend directly on beliefs; the latter two references take axiomatic approaches. Robin Pope has written extensively, over many years, about how conventional decision theory excludes various forms of belief-dependent motivation; Pope (2004) expounds her program and gives further references. Caplin & Leahy (2001) develop a model of “psychological expected utility” that admits belief-dependent motivation. However, these contributions mainly develop perspectives for settings with single decision-makers, and more will be needed to address games more generally.

Research on psychological games is still in its infancy. This is true for all facets of investigation: the development of basic classes of games and solution concepts, the investigation of decision-theoretic underpinning, tests of empirical (most likely experimental) validity, and finally applied theoretical work which uses psychological game theory to analyze various economic models. In each of these domains some work has been done which is indicative of the viability of the line of research, and there is good reason to be thrilled about the prospects for future research.

Martin Dufwenberg

[Suggested] See also: game theory, decision theory, emotions, social preferences, social rewards

Bibliography:

Bacharach, M., G. Guerra & D. Zizzo. 2001. Is Trust Self-Fulfilling? An Experimental Study. Mimeo.

Battigalli, P. & M. Dufwenberg. 2005. Dynamic Psychological Games. Mimeo.

Battigalli, P. & M. Siniscalchi. 1999. Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games. *Journal of Economic Theory* 88, 188-230.

Bell, D. 1985. Disappointment in Decision Making Under Uncertainty. *Operations Research* 33, 1-27.

Bernheim, B. D. 1994. A Theory of Conformity. *Journal of Political Economy* 102, 841-77.

Caplin, A. & J. Leahy. 2001. Psychological Expected Utility and Anticipatory Feelings. *Quarterly Journal of Economics* 116, 55-79.

Caplin, A. & J. Leahy. 2004. The Supply of Information by a Concerned Expert. *Economic Journal* 114, 487-505.

- Charness, G. & M. Dufwenberg. 2005. Promises & Partnership. Forthcoming in *Econometrica*.
- Dufwenberg, M. 1995. Time Consistent Wedlock with Endogenous Trust. In Doctoral Dissertation, *Economic Studies* 22, Uppsala University.
- Dufwenberg, M. 2002. Marital Investment, Time Consistency, and Emotions. *Journal of Economic Behavior & Organization* 48, 57-69.
- Dufwenberg, M. & U. Gneezy. 2000. Measuring Beliefs in an Experimental Lost Wallet Game. *Games & Economic Behavior* 30, 163-82.
- Dufwenberg, M. & G. Kirchsteiger. 2004. A Theory of Sequential Reciprocity. *Games & Economic Behavior* 47, 268-98.
- Dufwenberg, M. & M. Lundholm. 2000. Social Norms and Moral Hazard. *Economic Journal* 111, 506-25.
- Elster, J. 1989. Emotions and Economic Theory. *Journal of Economic Literature* 36, 47-74.
- Fehr, E. & S. Gächter. 2000. Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives* 14, 159-81.
- Geanakoplos, J., D. Pearce & E. Stacchetti. 1989. Psychological Games and Sequential Rationality. *Games & Economic Behavior* 1, 60–79.
- Gilboa, Y. & D. Schmeidler. 1988. Information Dependent Games: Can Common Sense Be Common Knowledge? *Economics Letters* 27, 215-21.
- Huang, P. & H.-M. Wu. 1994. More Order without More Law: A Theory of Social Norms and Organizational Cultures. *Journal of Law, Economics & Organization*, 10, 390-406.
- Karni, E. 1992. Utility Theory with Probability Dependent Outcome Valuation. *Journal of Economic Theory* 57, 111-24.
- Karni, E. & E. Schlee 1995. Utility Theory with Probability Dependent Outcome Valuation: Extensions and Applications. *Journal of Risk & Uncertainty* 10, 127-42.
- Loomes, G. & R. Sugden. 1986. Disappointment and Dynamic Consistency in Choice under Uncertainty. *Review of Economic Studies* 53, 271-82.
- Machina, M. 1981. “Rational” Decision Making versus “Rational” Decision Modeling. *Journal of Mathematical Psychology* 24, 163-175.
- Machina, M. 1989. Dynamic Consistency and Non-Expected Utility Models of Choice under Uncertainty. *Journal of Economic Literature* 27, 1622-1668.
- Pope, R. 2004. Biases from Omitted Risk Effects in Standard Gamble Utilities. *Journal of Health Economics* 23, 1029-50.

Rabin, M. 1993. Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83, 1281-1302.

Rabin, M. 1998. Psychology and Economics. *Journal of Economic Literature* 36, 11-46.

Sobel, J. 2005. Interdependent Preferences and Reciprocity. *Journal of Economic Literature* 43, 392-436.