Incorporating Belief-Dependent Motivations in Games

Pierpaolo Battigalli (Bocconi University)

Martin Dufwenberg (University of Arizona)

October 27, 2009

Abstract

Belief-dependent preferences allow to represent the impact of feelings and concerns about the feelings and opinions of others on decision making and strategic behavior. In Dynamic Psychological Games (JET, 2009, henceforth DPG), we put forward and analyze a substantial extension of the psychological games framework of Geanakoplos et al. (GEB, 1989) whereby the utilities of terminal nodes depend on hierarchies of conditional beliefs about strategies. Motivated by theoretical issues and applications, here we address three problems: (i) the derivation of utility functions defined on the extensive form (as those of DPG) from simpler and more intuitive functions that only depend on (material) consequences and beliefs about consequences; (ii) the dynamic inconsistency that arises when the derived utility function depends on beliefs about one's own behavior (a rather pervasive feature); (iii) the analysis of solution concepts that (unlike the sequential equilibrium of DPG) rely on the interpretation of observed behavior as the result of an intentional choice. Our framework provides the intellectual home for a rich variety of non-standard models of decision making and social interaction, and formally clarifies the distinction between anticipated and anticipatory feelings.

1 Introduction

Most (not all!) economic models assume that agents maximize their expected material payoff and hold equilibrium beliefs.

But subjects in the lab exhibit persistent and significant deviations from this self-interested, equilibrium behavior.

Furthermore, simple introspection suggests that agents are also affected by non self-interested motivations and that this affects strategic reasoning.

Models with *'other-regarding preferences'* help explain observed behavior in Dictator Game, Ultimatum G., Trust G., Gift Exchange G., Public Good G. and similar games

1. material payoffs of others matter: distribution-dependent preferences,

2. emotions and intentions matter: belief-dependent motivations.

(1) (distribution-dependent preferences) can be addressed by traditional game theory.

But experimental evidence (and, again, introspection) also support (2): theories of belief-dependent motivations (e.g. Dufwenberg & Gneezy 2000, Charness & Dufwenberg 2006, Dana *et al* 2006, Attanasi & Nagel 2007, Tadelis 2007). Our framework allows for (1), but it focuses on (2).

In "Dynamic Psychological Games" (DPG), we provide a new *framework* to deal with (*2*) (belief-dependent motivations) which is *outside traditional game theory*. The proposed framework also allows to (indeed, invites to) address the issue of relaxing the equilibrium assumption that players hold correct conjectures.

Loosely speaking: in a *psychological game* utility functions directly depend on (actions and) *beliefs*, including beliefs about the beliefs of others. Framework put forward by Geanakoplos, Pearce and Stacchetti (1989, henceforth GPS).

Our analysis concerns games with a *sequential* structure (dynamic games), e.g. Ultimatum, Trust, and Gift Exchange games. GPS' framework is *not* fully adequate for such games, because they only consider initial (pre-play) beliefs. In DPG we propose a new, more general framework where *revised beliefs about the beliefs of others* can play a role.

This is a follow up of DPG.

In DPG we consider (possibly infinite) *hierarchies of conditional beliefs* to model updated beliefs about the beliefs of others.

Here we mainly focus on *1st and 2nd order* beliefs.

On the other hand, here we:
- provide more structure for belief-dependent utility functions, deriving the abstract functional form of DPG from more elementary and intuitive utility functions,
- allow for players' uncertainty about their own behavior $\longrightarrow$ distinction *plans* vs *actual strategies*,
- focus on *dynamic inconsistency* and *perceived intentionality*.

ROADMAP

- Examples and motivation

- Trust Game and "guilt"

- Trust Game and "shame"

- From belief-dependent preferences to dynamic psychological games

- Game-form independent preferences

- Game-form dependent preferences

- Games with belief-dependent preferences

- Sequential rationality and solution concepts

- Dynamic (in)consistency

- Sequential equilibrium (SE)

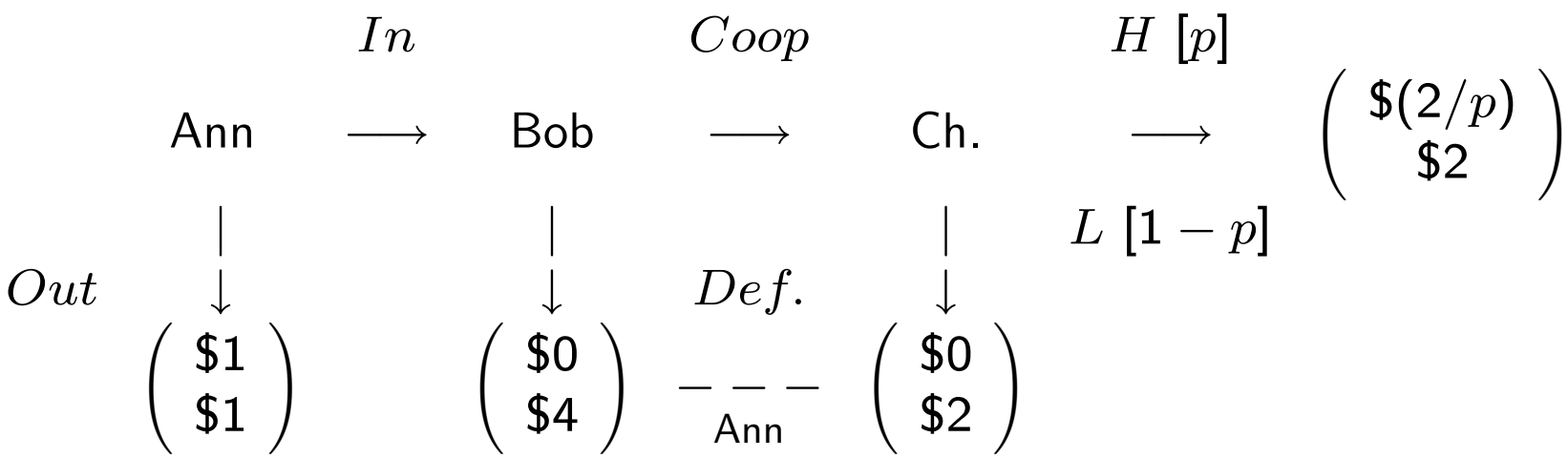
- Modifications of (SE): weak consistency and perceived intentionality

- Hints on other concepts related to learning and forward induction

- Summary/conclusions

2 Examples and motivation

Leading example: Trust Game form



Trust Game with material payoffs

Players observe *ex post* only material payoffs

(Why should we care? You will see!)

We will consider theories where the following beliefs play a crucial role:

$\alpha = \mathsf{Pr}_{Ann}[Coop \text{ if } In]$, the initial 1st-order belief of Ann

$\hat{\alpha} = \mathsf{Pr}_{Ann}[Coop|m_{Ann} = 0]$, a terminal (conditional) 1st order belief of Ann

$\beta = \mathsf{E}_{Bob}[\alpha|In]$ (a feature of) conditional 2nd-order belief of Bob

GPS' framework was used to model belief-dependent motivations:

in **traditional Game Theory**, payoff functions have the form

$$U_i = U_i(\text{actions})$$

actions=sequence of actions during play=*complete history*.

GPS' extension: psy-payoff functions

$$U_i = U_i(\text{beliefs}_i, \text{actions})$$

beliefs_i =*initial* (pre-play) beliefs of player i about strategies and beliefs (about beliefs) of *others*

Problem: conditional beliefs such as $\hat{\alpha}$ and β above are not considered by GPS (they are not part of the GPS language). Yet they are crucial for some applications, and theoretical interpretations of experimental findings.

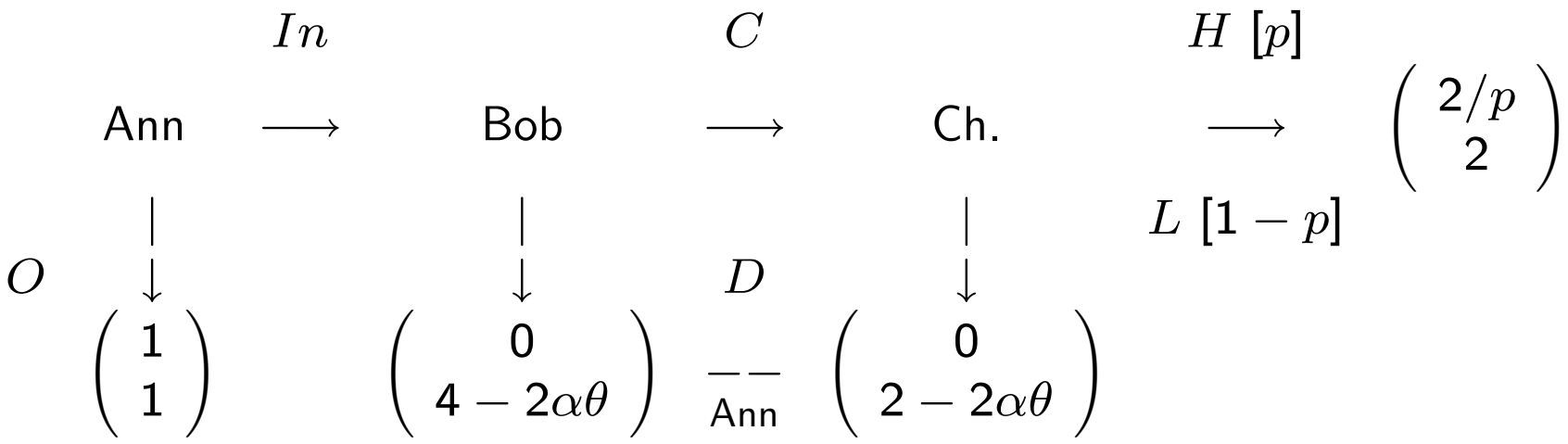
Our framework: we derive the following functional form

$$U_i(\text{cond.bel.}_i, \text{cond.bel.}_{-i}, \text{actions})$$

(we include beliefs about one's own future behavior), starting from more elementary utility functions.

Warning: While we appreciate work based on the revealed preference approach, we do not adopt it here.

Example: "guilt aversion", an easy to model motivation



Trust Game with guilt: dependence on co-player's 1st ord. belief

$\mathsf{E}_{Ann}[\widetilde{m}_{Ann}] = (1-\alpha)\cdot 0 + \alpha \cdot [p \cdot \frac{2}{p} + (1-p)\cdot 0] = 2\alpha$

2α=how much Ann would feel 'let down' if $m_{Ann} = 0$=difference btw expected and realized material payoff

2β=Bob's expectation of 2α, given *In* (conditional 2nd ord. belief)

θ=sensitivity of Bob to "guilt"

"primitive": $u_{Bob} = m_{Bob} - \theta \max\{0, \mathsf{E}_{Ann}[\widetilde{m}_{Ann}] - m_{Ann}\}$

derived: $U_{Bob} = \mathbf{m}_{Bob}(z) - \theta \max\{0, \mathsf{E}_{Ann}[\widetilde{m}_{Ann}] - \mathbf{m}_{Ann}(z)\}$

Bob would (weakly) prefer *Coop.* after *In* iff

$$\begin{aligned} 2-(1-p)\cdot 2\beta\theta &\geq 4-2\theta\beta \\ \theta &\geq \frac{1}{\beta p} \end{aligned}$$

Equilibria (intuitive analysis):

In, Coop., $\alpha=\beta=1$ if θ is high enough ($\theta>1/p$) is an equil.

But also *Out*, $\alpha=0,\beta=0$ is an equilibrium (for all θ).

(Attanasi & Nagel show that $\theta>1$ is quite realistic.)

NOTE: such multiplicity of eq. is ruled out by standard game theory, if complete information (=common knowledge of the game) is assumed, because PI games cannot have multiple isolated equilibria.

Equivalent representation: Ann's (initial) 1st order beliefs, Bob's own terminal 2nd order beliefs

$$\begin{aligned} U_B &= \mathbf{m}_B(z) - \theta \mathsf{E}_B\left[\max\{0, \mathsf{E}_A[\widetilde{m}_A] - \widetilde{m}_A\} | \widetilde{m}_B = \mathbf{m}_B(z)\right] \\ &= U_B(z, \text{cond. 2nd ord. beliefs of Bob}) \end{aligned}$$

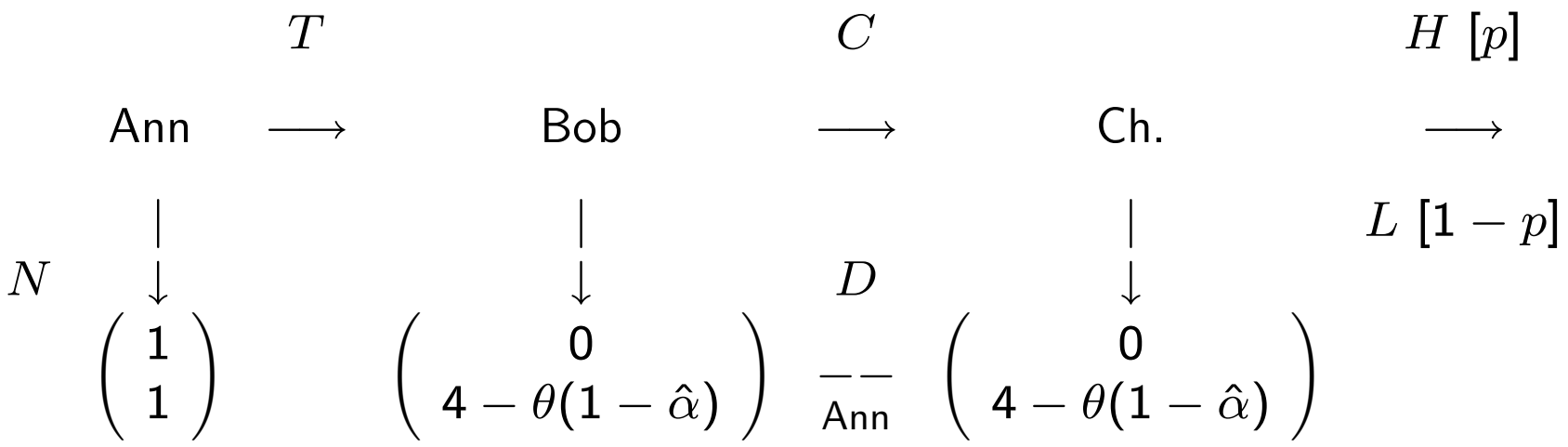
$$\begin{array}{ccccccc} & & In & & C & & H\ [p] \\ & \text{Ann} & \longrightarrow & \text{Bob} & \longrightarrow & \text{Ch.} & \longrightarrow \\ & \Big\downarrow & & \Big\downarrow & & \Big\downarrow & L\ [1-p] \\ O & \begin{pmatrix} 1 \\ 1 \end{pmatrix} & & \begin{pmatrix} 0 \\ 4-2\beta\theta \end{pmatrix} & \underset{\text{Ann}}{D} & \begin{pmatrix} 0 \\ 2-2(1-p)\beta\theta \end{pmatrix} & \end{array} \quad \left(2 - \right.$$

Trust Game w/ guilt: dependence on own terminal higher order beliefs

Example: simple "Shame", Bob dislikes that Ann thinks he has defected,

$$u_{Bob} = m_{Bob} - \theta(1 - \hat{\alpha}) \qquad (\hat{\alpha} = \mathsf{Pr}_{Ann}[Coop|m_{Ann} = 0])$$

(easy to model in Trust Game; but generally applicable functional form not easy)



Trust Game w/ shame: dependence on co-player's term. 1st ord. belief

Let

$$\hat{\beta}_D = \mathsf{E}_{Bob}\left[\Pr_{Ann}[Coop|m_A = 0]\middle|(In, D)\right] = \mathsf{E}_{Bob}\left[\frac{\alpha - \alpha p}{1 - \alpha p}\middle|(In, D)\right]$$

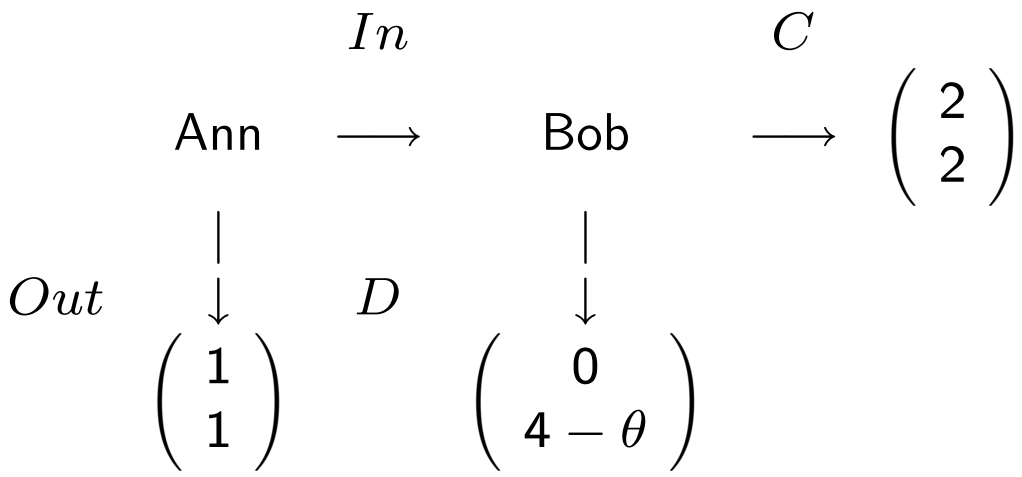
$$\hat{\beta}_C = \mathsf{E}_{Bob}\left[\Pr_{Ann}[Coop|m_A = 2/p]\middle|(In, C)\right] = p + (1-p)\hat{\beta}_D$$

Bob's decision depends on $\hat{\beta}_D$ and $\hat{\beta}_C$, the game looks to him like this

$$\begin{array}{ccccc} & & In & & C \\ & \text{Ann} & \longrightarrow & \text{Bob} & \longrightarrow \begin{pmatrix} 2 \\ 2 - \theta(1 - \hat{\beta}_C) \end{pmatrix} \\ & \Big\downarrow & & \Big\downarrow & \\ Out & D & & \begin{pmatrix} 0 \\ 4 - \theta(1 - \hat{\beta}_D) \end{pmatrix} & \\ \begin{pmatrix} 1 \\ 1 \end{pmatrix} & & & & \end{array}$$

Reduced-Form, Trust Game w/ Shame from Bob's point of view

Now suppose that players (in particular, Ann) have *perfect ex post information*. Then the game looks to Bob like this



Reduced-Form, Trust Game w/ Shame, *perfect ex post information*

Bob is more likely to Cooperate under ex post perfect information, anticipating this *Ann trusts Bob more* (more likely to play *In*).

Experimental evidence supports this result on the impact of the ex post information structure (Tadelis 2007, see also Dana et al. 2006).

According to standard game theory preferences depend only on actions and random events, and this implies that only the information the players have when they are active may be relevant. Thus, *contrary to experimental evidence, standard game theory rules out the impact of ex post information.* This shows that *observed phenomena explained with belief-dependent preferences cannot be explained by standard game theory*, even allowing for incomplete information (despite the opposite claims by some "orthodox" theorists).

[The previous conclusion must be qualified: it relies on the assumption that the relevant game form is the one specified in the lab, i.e. that interaction in the lab is effectively isolated from postexperimental interactions outside the lab. This assumption is disputable when self-perception (e.g. self-esteem) matters: it can be argued that ex post information may affect self-perception in the post-experiment life, and this may affect subjects' behavior in the experiment. But we think this argument does not apply to the previous example.]

3 From belief-dependent motivations to games

3.1 Game form and beliefs

Finite extensive game form: $\Gamma = \langle N, X, (H_i, M_i, \mathbf{m}_i)_{i \in N} \rangle$

- *material* consequences of terminal nodes $z \in Z$: $m_i = \mathbf{m}_i(z) \in M_i$ (e.g., $M_i \subseteq \mathbb{R}$, "*money*")

- H_i partitions X, player i's *information* $h = H_i(x) \in H_i$ specified *at every node* x (even if i is *not* active), including root x^0 and terminal nodes z; assume *perfect recall*; $h^0 := \{x^0\} \in H_i$ for each i; $\hat{H}_i \subset H_i$ information sets where i is *active*

- *chance*=fictitious player with exogenous behavior strategy σ_c

- derive *outcome function* $z = \mathbf{z}(s)$ (s=strategy profile)

- *behavior strategies* $\sigma_i = (\sigma_i(\cdot|h))_{h \in \hat{H}_i}$, derive $\mathsf{Pr}_{\sigma_i}(s_i|h)$

Hierarchical conditional beliefs (general analysis in Battigalli-Siniscalchi, 1999)

- 1st-order cond. belief system $\alpha_i = (\alpha_i(\cdot|h))_{h \in H_i}$, $\alpha_i(\cdot|h) \in \mathbf{\Delta}(S)$

- 2nd order: $\beta_i = (\beta_i(h))_{h \in H_i}$, $\beta_i(h)$=point belief (just for simplicity) of i given h about $\alpha_{-i} = (\alpha_j)_{j \neq i}$

Assume it is true and "transparent" that Bayes rule holds.
The set of 1st order cond. belief systems of i satisfying Bayes rule is denoted $\mathbf{\Delta}^{H_i}(S)$ (a subset of $[\mathbf{\Delta}(S)]^{H_i}$).

3.2 Simple "game-form free" belief dependent motivations

2 players: Ann (pl. A) and Bob (pl. B). In our prose we mostly take Ann's perspective. Likewise statements hold for Bob

Periods $t = 1, 2, ..., T$ (T "endogenous"),

$M = M_A \times M_B$ (set of collective *m*aterial consequences)

$\mu_A^0 \in \Delta(M)$ initial belief of Ann about consequences

$\mu_A^t \in \Delta(M)$ end-of-period t belief of Ann

Assumption: the utility of consequence m for Ann *depends on the temporal sequence of beliefs* experienced by Ann *and* Bob $\Rightarrow$ "primitive" psychological utility function

$$u_A((\mu_A^0, \mu_B^0), ..., (\mu_A^T, \mu_B^T), m)$$

$$u_A : (\mathbf{\Delta}(M) \times \mathbf{\Delta}(M))^* \times M \rightarrow \mathbb{R}$$

[$Y^* =$ set of finite sequences of elements from domain Y]

Motivations:

- μ_A^t may trigger anticipatory feelings of Ann with negative or positive valence, such as *anxiety* or excitement, that affect Ann's period-t utility (Caplin & Leahy)

- Some time-t feelings, such as *disappointment*, may depend on earlier beliefs μ_A^k $(k < t)$.

- The *anticipation of such feelings* (e.g. terminal feeling or interim anticipatory feelings) may affect behavior. This can be represented as maximization of an intertemporal utility function as above.

- Ann may care for the feelings of Bob (e.g., guilt, shame, concern for other's anxiety).

Now *fix a game form* Γ. Recall:
$m_i = \mathbf{m}_i(z) = i$'s material consequence,
$z = \mathbf{z}(s)$= terminal history induced by s.
Let $\mathbf{m}(z) := (\mathbf{m}_A(z), \mathbf{m}_B(z))$.

Derive $\mu_A \in \mathbf{\Delta}(M)$ at $h \in H_A$ from $\alpha_A = (\alpha_A(\cdot|h))_{h \in H_A}$

$$\mu_{\alpha_A}(m|h) = \sum_{s:\mathbf{m}(\mathbf{z}(s))=m} \alpha_A(s|h) \ (h \in H_A)$$

Fix complete path $(x^0, x^1, ..., x^T = z)$, then

$$U_A(\alpha_A, \alpha_B, z) = u_A\left(\left(\mu_{\alpha_A}(\cdot|H_A(x^t)), \mu_{\alpha_B}(\cdot|H_B(x^t))\right)_{t=0}^{T}, \mathbf{m}(z)\right)$$

In words, $U_A(\alpha_A, \alpha_B, z)$ is Ann's utility of the temporal sequences of beliefs and the collective consequence induced by z given the 1st-order cond. belief systems of Ann and Bob, α_A, α_B.

Examples of "primitive" utility functions

- *Anxiety*: $u_A = m_A + \sum_{t=0}^{T-1} \left\{ \theta_A^{E,t} E_{\mu_A^t}[\widetilde{m}_A] - \theta_A^{V,t} \mathsf{Var}_{\mu_A^t}[\widetilde{m}_A] \right\}$ (Caplin & Leahy QJE)

- Simple "final" *disappointment*: $u_A = m_A - \theta_A \max\{\mathsf{E}_{\mu_A^0}[\widetilde{m}_A] - m_A, 0\}$

...and concerns for such feelings e.g.

- "guilt": $u_A = m_A - \theta_A \max\{\mathsf{E}_{\mu_B^0}[\widetilde{m}_B] - m_B, 0\}$ (Batt.-Duf. AER)

- concern for other's anxiety:
$u_A = m_A + \sum_{t=0}^{T-1} \left\{ \theta_A^{E,t} E_{\mu_B^t}[\widetilde{m}_B] - \theta_A^{t} \mathsf{Var}_{\mu_B^t}[\widetilde{m}_B] \right\}$ (Caplin & Leahy EJ)

Extension: self-esteem and concern for the opinion of others

Suppose we can capture a feature of Ann's ability or personality with a parameter θ_A *not commonly known* (e.g. Bob does not perfectly know Ann's preferences, Bob and/or Ann do not perfectly know the ability of Ann).

Ann's feelings (with positive or negative valence) may be affected be her beliefs about θ_A (self-esteem).

Ann may care about Bob's opinion of her, i.e. Bob's beliefs about θ_A.

We can capture these considerations with the following extension:

$$u_A((\mu_A^0, \mu_B^0), ..., (\mu_A^T, \mu_B^T), m)$$
$$u_A \ : \ (\Delta(\Theta \times M) \times \Delta(\Theta \times M))^* \times M \to \mathbb{R}$$

where $\Theta = \Theta_A \times \Theta_B$. As special cases we get:

Ex post self-esteem: u_A depends only on $\text{marg}_{\Theta_A}\mu_A^T$

Concern for the ex post opinion of others: u_A depends only on $\text{marg}_{\Theta_A}\mu_B^T$

With this extension we cover most models in the literature in which preferences can be expressed with no reference to an extensive game form.

Now fix and extensive game form Γ. Let $\alpha_i = \left(\alpha_i(\cdot|h)_{h\in H_i}\right)$, $\alpha_i(\cdot|h) \in \Delta(\Theta \times S)$ (and i always assigns probability one to what she knows about θ). With a procedure similar to the one we explained above we get

$$U_A(\theta, \alpha_A, \alpha_B, z)$$

3.3 Game-form & belief dependent motivations

Reciprocity, regret, concern for others' regret

Relevance of the game form that determines the strategy space (causal structure, available options).

"Primitive" utility function

$$u_A : (\Delta(S) \times \Delta(S))^* \times M \to \mathbb{R}$$

Fix complete path $(x^0, x^1, ..., x^T = z)$, then

$$U_A(\alpha_A, \alpha_B, z) = u_A\left(\left(\alpha_A(\cdot|H_A(x^t)), \alpha_B(\cdot|H_B(x^t))\right)_{t=0}^{T}, \mathbf{m}(z)\right).$$

Examples of game-form dependent preferences

Regret. Extent of Ann's regret when she gets m_A and her *terminal* belief about Bob's strategy is $\alpha^T_{A,B} \in \Delta(S_B)$:

$$R_A(\alpha^T_{A,B}, m_A) = \max_{s_A} \sum_{s_B} \mathbf{m}_A(\mathbf{z}(s_A, s_B))\alpha^T_{A,B}(s_B) - m_A,$$

"Primitive" and derived utility:

$$u_A((\alpha^t_A, \alpha^t_B)^T_{t=0}, m) = m_A - f(R_A(\alpha^T_{A,B}, m_A)),\ (f(0) = 0, f' > 0)$$

$$U_A(\alpha_A, \alpha_B, z) = \mathbf{m}_A(z) - f(R_A(\text{marg}_{S_B}\alpha_A(\cdot|H_A(z)), \mathbf{m}_A(z)))$$

Concern for others' regret. Ann may be concern for the regret of Bob (e.g. her son). Thus regret may be relevant even if it does not directly affect the behavior of the regretting person.

"Primitive" and derived utility:

$$u_A((\alpha_A^t, \alpha_B^t)_{t=0}^T, m) = m_A - f(R_B(\alpha_{B,A}^T, m_B)),\ (f(0) = 0, f' > 0)$$

$$U_A(\alpha_A, \alpha_B, z) = \mathbf{m}_A(z) - f(R_B(\text{marg}_{S_A}\alpha_B(\cdot|H_B(z)), \mathbf{m}_B(z)))$$

Reciprocity (cf., Rabin, Duf. & Kirchsteiger, Falk & Fischbacher; a bit different here)

Define the *kindness of Bob toward Ann* as a function of Bob's beliefs about strategies, $\alpha_B \in \Delta(S)$

$$K_{B,A}(\alpha_B) = \mathsf{E}_{\alpha_B}[\widetilde{m}_A] - m^e_A(\alpha_B),$$

where $m^e_A(\alpha_B)$ is the "equitable" payoff for Ann according to Bob's beliefs, for example

$$m^e_A(\alpha_B) = \frac{1}{2}\left(\max_{s_B} \mathsf{E}_{s_B,\alpha_{B,A}}[\widetilde{m}_A] + \min_{s_B} \mathsf{E}_{s_B,\alpha_{B,A}}[\widetilde{m}_A]\right) \quad [\alpha_{B,A} \in \Delta(S_A)]$$

"Primitive" and derived utility of Ann:

$$u_A((\alpha^t_A, \alpha^t_B)^T_{t=0}, m) = m_A + \theta_A K_{B,A}(\alpha^0_B) m_B,\ (\theta_B \geq 0).$$

$$U_A(\alpha_A, \alpha_B, z) = \mathbf{m}_A(z) + \theta_A K_{B,A}(\alpha_B(\cdot|h^0))\mathbf{m}_B(z).$$

3.4 Games with belief-dependent preferences

Adding the derived utility functions

$$U_i : \Delta^{H_A}(S) \times \Delta^{H_B}(S) \times Z \to \mathbb{R}$$

to the game form Γ we obtain a (dynamic) *game with belief-dependent preferences*

$$\langle \Gamma, U_A, U_B \rangle$$

called "dynamic psychological game" in DPG.

[In DPG we consider more general functional forms allowing dependence on higher order beliefs, but we mainly focus on the case where U_A is independent of Ann's plan, i.e. her belief about her own strategy.]

4 Solution concepts

To keep things simple assume that there are two players, A and B [plus chance: add it as 3rd player with known randomized strategy σ_c]
Slight **abuse of notation**: if $\mu \in \Delta(Y)$ and $\mu(\{y\}) = 1$, write $\mu = y \in Y$ [hence $Y \subset \Delta(Y)$]

- $\alpha_A \in \Delta^{H_A}(S)$ [system of 1st order cond. beliefs of Ann] satisfies *independence across agents*, that is, it can be derived from a profile of behavior strategies $\sigma_{\alpha_A} = \left(\sigma_{A,\alpha_A}, \sigma_{B,\alpha_A}\right)$ using Kuhn's transformation [$\sigma_{i,\alpha_A}(a_i|h)$ $(i = A, B)$ is the cond.prob. of a_i given h implied by α_A]

- β_A is given by *point beliefs* about α_B: let $\beta_A(h)$ denote the second-order point belief of Ann at $h \in H_A$ [formally, $\beta_A(h) \in \Delta^{H_B}(S)$]

Likewise for Bob

The system of beliefs of Ann is an array $(\alpha_A, \beta_A) = (\alpha_A(\cdot|h), \beta_A(h))_{h \in H_A}$ where $\alpha_A \in \Delta^{H_A}(S)$, and $\beta_A = (\beta_A(h))_{h \in H_A}$ is s.t. for all $\hat{h}, h \in H_A$ with $\hat{h} \prec h$

- $\beta_A(h) \in \Delta^{H_B}(S)$,
- if $\alpha_{A,B}(S_B(h)|\hat{h}) > 0$ then $\beta_A(h) = \beta_A(\hat{h})$.

In words, Ann may change her point belief about the first-order beliefs of Bob only if she is surprised by the behavior of Bob. Call this *simple system of beliefs*

4.1 Dynamic (in)consistency

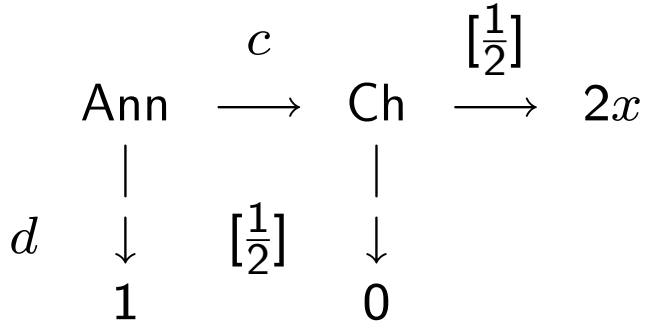
Let $\alpha_{A,A} = \left(\text{marg}_{S_A}\alpha_A(\cdot|h)\right)_{h\in H_A}$, $\alpha_{A,B} = \left(\text{marg}_{S_B}\alpha_A(\cdot|h)\right)_{h\in H_A}$.
We interpret Ann's beliefs about her own strategy, $\alpha_{A,A}$, as her *plan*. First-order belief α_A captures Ann's *intentions*, i.e. what she plans to do and expects to achieve.

Problem 1: When U_A depends on $\alpha_{A,A}$, Ann's beliefs about her own strategy (e.g. disappointment or anxiety aversion), we may have *dynamic inconsistency* $\Rightarrow$ use def. of "sequential best response" as *no incentive for one-shot deviations:*

$$\forall h \in \hat{H}_i,\ \text{supp}\sigma_{i,\alpha_i}(\cdot|h) \subseteq \arg\max_{a_i\in A_i(h)} \mathsf{E}_{\alpha_i,\beta_i}[U_i|h,a_i] \qquad \text{(BR)}$$

NOTE, we allow a player to be uncertain about her own strategy and require that there are no incentives to deviate to zero-prob. actions. Such uncertainty may be necessary for existence: in some psy-games where U_i depends on $\alpha_{i,i}$ no "pure" $\alpha_{i,i}$ satisfies the seq. rationality property above.

Example: aversion to "excess disappointment" may prevent pure dynamically consistent plans



One person game form with chance $(1 < x < 2)$

Ann is averse to disappointment in excess of $k > 0$:

$$u_A(\mu_A^0, m) = m - \theta \max\left\{\left(\mathsf{E}_{\mu_A^0}[\widetilde{m}] - m - k\right), 0\right\}$$

Here, let $k = x - 1$ to simplify the algebra.

If the initial plan $(\alpha^0_{A,A})$ is d (down):
$\mathsf{E}_{\alpha^0_A}[\widetilde{m}] = 1$, $U_A(\alpha^0_A,(d)) = 1$, $U_A(\alpha^0_A,(c,H)) = 2x$, $U_A(\alpha^0_A,(c,L)) = -\theta(2-x)$,
switch $d \to c$ iff $\mathsf{E}_{\alpha_A}[U_A|d] < \mathsf{E}_{\alpha_A}[U_A|c]$ iff $1 < x - \frac{1}{2}\theta(2-x)$ iff $\theta < \frac{2(x-1)}{2-x}$.

If the initial plan $(\alpha^0_{A,A})$ is c (continue to chance move) :
$\mathsf{E}_{\alpha^0}[\widetilde{m}] = x$, $U_A(\alpha^0_A,d) = 1$, $U_A(\alpha^0_A,(c,H)) = 2x$,
$U_A(\alpha^0_A,(c,L)) = -\theta$,
switch $c \to d$ iff $\mathsf{E}_{\alpha_A}[U_A|d] > \mathsf{E}_{\alpha_A}[U_A|c]$ iff $1 > x - \frac{1}{2}\theta$ iff $\theta > 2(x-1)$.

There is no pure dynamically consistent plan iff $2(x-1) < \theta < \frac{2(x-1)}{2-x}$,
e.g. $x = \theta = \frac{3}{2}$.

Some interesting belief-dependent motivations (e.g. concern for others' feelings, regret, some forms of reciprocity) yield a U_A that does not depend on A's plan of action:

Remark. *If U_A does not depend on $\alpha_{A,A}$ then Ann is dynamically consistent, that is, condition (BR) is equivalent to*

$$\forall h \in \hat{H}_A, \text{supp}\alpha_{A,A}(\cdot|h) \subseteq \arg \max_{s_A \in S_A(h)} \mathsf{E}_{s_A,\alpha_{A,B},\beta_A}[U_A|h].$$

4.2 Sequential equilibrium

DEF. 1 A simple system of beliefs $(\alpha_A, \beta_A, \alpha_B, \beta_B)$ is a *sequential equilibrium* (cf. Kreps & Wilson) if for each i, each $h \in H_i$,

$$\forall h \in H_i,\ \alpha_i = \alpha_{-i} = \beta_i(h), \tag{1}$$

$$\forall h \in \hat{H}_i,\ \text{supp}\sigma_{i,\alpha_i}(\cdot|h) \subseteq \arg \max_{a_i \in A_i(h)} \mathsf{E}_{\alpha_i,\beta_i}[U_i|h, a_i]$$

- (CONS) says that first-order beliefs of different players agree, and second-order conditional beliefs are always correct, implying that i *cannot change his mind about the (first-order) beliefs of the co-player*. [With chance moves or more players, we must add a further condition to ensure consistency of assessments.]
- The 2nd requirement is the "local" best response property.

This is a "natural" and relatively simple extension of the SE concept of K.W.

Remark: *Suppose that U_i does not depend on $\alpha_{i,i}$ $(i = A, B)$. Then a simple system of beliefs $(\alpha_A, \beta_A, \alpha_B, \beta_B)$ is a sequential equilibrium if and only if for each i,*

$$\begin{aligned} \forall h &\in H_i,\ \alpha_i = \alpha_{-i} = \beta_i(h), \\ \forall h &\in \hat{H}_i,\ \mathrm{supp}\alpha_{i,i}(\cdot|h) \subseteq \arg \max_{s_i \in S_i(h)} \mathsf{E}_{\alpha_{i,-i},\beta_i}[U_A|h] \end{aligned}$$

Now suppose that U_i does not depend on $\alpha_{i,i}$ *nor* on $\alpha_{-i,-i}$ (a strong form of own-plan independence). Then the above result implies that we may interpret a sequential equilibrium as a situation where each player is certain of her own strategy and the randomized strategy of i represents the 1st order beliefs of $-i$ about i.

By quite standard "trembling hand" arguments, one can show that a sequential equilibrium always exist [cf. DPG]

Theorem: *Every (finite) game with belief-dependent preferences has a sequential equilibrium.*

4.3 A more satisfactory equilibrium concept

Problem 2: Sequential equilibrium *à la* Kreps and Wilson is based on *correct initial beliefs* (including 2nd order beliefs: $\beta_i(h^0) = \alpha_{-i}$) and a *trembling hand* interpretation of deviations $\Rightarrow i$ never changes his beliefs about α_{-i} ($\forall h \in H_i$, $\beta_i(h) = \beta_i(h^0) = \alpha_{-i}$).

This is problematic in general: Bob's beliefs ($\beta_B(h)$) about Ann's intentions (α_A) are independent of Ann's behavior. It is *even more problematic in psychological games where players' intentions are key.*

Example: sequential equilibrium and reciprocity in the Trust Game

$m^e_B(\alpha_{A,B}) = \frac{1}{2}[1 + 2 \times \alpha_{A,B}(Coop) + 4 \times \alpha_{A,B}(Def.)] \in [\frac{3}{2}, \frac{5}{2}]$
="equitable payoff of Bob" (in Ann's eyes)

$K_{A,B}(\alpha_A) = E_{\alpha_A}[\mathbf{m}_B] - m^e_B(\alpha_{A,B})$

$K_{A,B}(Out, \alpha_{A,B}) = 1 - m^e_B(\alpha_{A,B}) < 0$

$K_{A,B}(In, \alpha_{A,B}) = 2 \times \alpha_{A,B}(Coop) + 4 \times \alpha_{A,B}(Def) - m^e_B(\alpha_{A,B})$
$= \frac{1}{2} + \alpha_{A,B}(Def) > 0.$

In is unquestionably a kind action, if it is intentional! If Bob believes, even if surprised, that Ann's choice was intentional, and he is sensitive to Ann's kindness (high θ_B), he should reciprocate and Coop.

But (Out, Def) *is a sequential eq. for every* θ_B. Why? Because with $\alpha_A = (Out, Def)$, $K_{A,B}(\alpha_A) < 0$, after In Bob still believes that Ann's intentions are given by $\alpha_A = (Out, Def)$ (Ann planned Out and chose In "by mistake"), hence Bob still perceives Ann as unkind!

We want to allow Bob to change his beliefs about Ann's intentions (captured by α_A) when he is surprised by Ann's choice, for example because he believes that her observed choice was part of her plan, and/or because he "rationalizes" Ann's choice (forward induction reasoning).

We start with a weaker notion of equilibrium than SE, by weakening the consistency requirement (CONS).

DEF. 2 A simple system of beliefs $(\alpha_A, \beta_A, \alpha_B, \beta_B)$ is a *weakly consistent perfect Bayesian equilibrium* if for each i,

$$\alpha_i(\cdot|h^0) = \alpha_{-i}(\cdot|h^0),\ \beta_i(h^0) = \alpha_{-i} \qquad \text{(Weak CONS)}$$

$$\forall h \in \hat{H}_i,\ \text{supp}\sigma_{i,\alpha_i}(\cdot|h) \subseteq \arg \max_{a_i \in A_i(h)} \mathsf{E}_{\alpha_i,\beta_i}[U_i|h, a_i]$$

Intuition (weak CONS.) *Initial 1st order beliefs agree, and players* **start** *with correct 2nd order beliefs* [recall that by def. of belief system, Bob keeps the same beliefs about Ann's 1st order beliefs as long as he is not surprised by Ann's choices: $\alpha_{B,A}(S_A(h)|h^0) > 0$].

Let $\beta_i^0(h)$ be i's *point belief at* h *about* $\alpha_{-i}(\cdot|h^0)$, thus $\beta_i^0(h) \in \Delta(S)$ and it makes sense to write $\text{marg}_{S_j}\beta_i^0(h) \in \Delta(S_j)$ for the belief of i (at h) on the belief of $-i$ on the behavior of j ($j = i, -i$). In particular, $\text{marg}_{S_{-i}}\beta_i^0(h)$ is i's belief at h about $-i$'s plan.

DEF. 3 A weakly consistent PBE $(\alpha_A, \beta_A, \alpha_B, \beta_B)$ satisfies *perceived intentionality* if for each i

$$\forall h \in H_i,\ \alpha_{i,-i}(S_{-i}(h)|h) = 0 \Rightarrow \text{suppmarg}_{S_{-i}}\beta_i^0(h) \subseteq S_{-i}(h).$$

Intuition: Whenever Bob observes an unexpected action by Ann, he *believes that* Ann planned to take that action, i.e. that *it was not chosen by mistake*.

Example: reciprocity in the Trust Game revisited

Suppose that Bob is sufficiently reciprocal ($\theta_B > 2$), then the only weakly cons. PBE that satisfies perceived intentionality is $(In, Coop)$:

- recall $U_B(\alpha_A, \alpha_B, z) = \mathbf{m}_B(z) + \theta_B K_{A,B}(\alpha_A(\cdot|h^0))\mathbf{m}_A(z)$,
$K_{A,B}(In, \alpha_{A,B}) = \frac{1}{2} + \alpha_{A,B}(Def)$

- assuming perceived intentionality:
$U_B(\alpha_{A,B}, (In, Def)) = 4$,
$U_B(\alpha_{A,B}, (In, Coop)) = 2 + \theta_B \times \left(\frac{1}{2} + \alpha_{A,B}(Def)\right) \times 2$

$\Rightarrow$ Bob prefers $Coop$ whatever $\alpha_{A,B}$ if $\theta_B > 2$

4.4 Rationalizability and self-confirming equilibrium

Problem 3: The assumption of correct (initial) beliefs is problematic in general, even more so in psychological games where higher-order beliefs are key $\Rightarrow$ apply notions of *rationalizability* (as in DPG) and/or *self-confirming equilibrium*, for example

- *simple self-confirming equilibrium:* players are (1) "sequentially rational" and (2) their beliefs are confirmed (beliefs about observables are correct), or

- *rationalizable self-confirming equilibrium:* (1) and (2) and initial common belief in (1)&(2), or

- *rationalizable self-confirming equilibrium with F.I.:* (1) and (2) and common *strong* belief in (1)&(2) (cf. Battigalli-Siniscalchi, 2002)

A correct analysis of these issues makes it necessary to go beyond the simplifying assumptions made above and consider more general hierarchies of conditional beliefs:

- use beliefs above the 2nd order to model forward induction and common belief

- do not assume point higher order beliefs, e.g. to model uncertainty about the intentions of a co-player

DPG does this under the assumption that players "know" their actual contingent behavior and there is common belief of this. But such assumption works well only under own-plan independence, which rules out interesting belief-dependent motivations.

5 Summary

- Higher-order conditional beliefs are necessary to model interesting belief-dependent motivations in dynamic games: go beyond Geanakoplos et al. (use work of Battigalli & Siniscalchi on hierarchies of conditional beliefs)

- Belief-dependent preferences in games yield qualitative predictions (in particular, comparative statics) that cannot be obtained within standard game theory, not even allowing for incomplete information

- Psychological utility functions *à la* Batt.-Duf. DPG, that depend on (terminal nodes and) systems of conditional beliefs, can be derived from more primitive utility functions that depend on temporal sequences of beliefs (own and others') about material consequences (game-form free), or about strategies (game-form dependent)

- Many interesting belief-dependent preferences can be simply modeled by assuming that psychological utility depends only on *own* and *others' first-order beliefs*

- Traditional notions of equilibrium, such as sequential equilibrium, are questionable in general, but even more so when extended to psychological games: look for intention-based notions of equilibrium, relax the assumption of correct beliefs

- We propose a new, very flexible framework that already encompasses a host of exciting applications (e.g. sequential reciprocity, guilt, anxiety, shame, disappointment, regret) and hopefully will allow and prompt many other (applications of such belief-dependent preferences to economic models, new forms of belief-dependent preferences)

References

[1] G. Attanasi and R. Nagel, A survey of psychological games: theoretical findings and experimental evidence, in: A. Innocenti and P. Sbriglia (Eds.), *Games, Rationality and Behaviour. Essays on Behavioural Game Theory and Experiments*, Palgrave McMillan, Houndmills, 2007, pp 204-232.

[2] P. Battigalli and M. Dufwenberg, Dynamic psychological games, *J. Econ. Theory* 144 (2009), 1-35.**[DPG]**

[3] P. Battigalli and M. Dufwenberg, Guilt in Games, *Amer. Econ. Rev. (P&P)* 97 (2007), 170-176.

[4] P. Battigalli, P. and M. Siniscalchi, Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games, *J. Econ. Theory* 88 (1999), 188-230.

[5] P. Battigalli, P. and M. Siniscalchi, Strong Belief and Forward Induction Reasoning, *J. Econ. Theory* 106 (2002), 356-391.

[6] A. Caplin and J. Leahy, Psychological expected utility theory and anticipatory feelings, *Quart. J. Econ.* 116 (2001), 55-79.

[7] A. Caplin and J. Leahy, The supply of information by a concerned expert, *Econ. J.* 114 (2004), 487-505.

[8] G. Charness and M. Dufwenberg, Promises and Partnership, *Econometrica* 74 (2006), 1579-1601.

[9] J. Dana, D.M. Cain, R. Dawes, What you don't know won't hurt me: Costly (but quiet) exit in dictator games, *Organizational Behavior and Human Decision Processes* 100 (2006), 193-201.

[10] M. Dufwenberg and U. Gneezy, Measuring beliefs in an experimental lost wallet game, *Games Econ. Behav.* 30 (2000), 163-182.

[11] M. Dufwenberg and G. Kirchsteiger, A theory of sequential reciprocity, *Games Econ. Behav.* 47 (2004), 268-298.

[12] A. Falk and U. Fischbacher, A theory of reciprocity, *Games Econ. Behav.* 54 (2006), 293-315.

[13] J. Geanakoplos, D. Pearce and E. Stacchetti, Psychological games and sequential rationality, *Games Econ. Behav.* 1 (1989), 60-79.**[GPS]**

[14] B. Kőszegi, Emotional agency, *Quart. J. Econ.* 12 (2006), 121-156.

[15] B. Kőszegi and M. Rabin, A model of reference-dependent preferences, *Quart. J. Econ.* 121, 1133-1166 (2006).

[16] B. Kőszegi and M. Rabin, Reference-dependent consumption plans, *Amer. Econ. Rev.* 99 (2007), 909-936.

[17] M. Rabin, Incorporating fairness into game theory and economics, *Amer. Econ. Rev.* 83 (1993), 1281-1302.

[18] S. Tadelis, The power of shame and the rationality of trust, mimeo, UC Berkeley, 2007.