Logic and Logical Philosophy
Online First Articles (2026)
DOI: 10.12775/LLP.2026.005

Eleonora Cresto

Dynamic Logic for Ungrounded Payoffs

Abstract. Higher-order likes and desires sometimes lead agents to have ungrounded or paradoxical preferences. This situation is particularly problematic in the context of games. If payoffs are interdependent, the overall assessment of particular courses of action becomes ungrounded; in such cases, the game's matrix is radically underdetermined. Paradigmatic examples of this phenomenon occur when players are 'perfect lovers' or 'perfect haters', in a sense to be explained. In this paper, I use a dynamic doxastic logic to mimic the search for a suitable matrix. Upgrades are triggered by conjectures about other players' utilities, which can, in turn, be based on behavioural or verbal cues. We can prove that, under certain conditions, pairs of agents with paradoxical preferences eventually come to believe they cannot interact in a game. As a result, I hope to provide a better understanding of game-theoretic ungroundedness and, more generally, of the structure of higher-order preferences and desires.

Keywords: preference logic; game theory; higher-order preferences; belief change; interdependent payoffs; paradoxicality

Introduction

Anna wished Bob and Tom wouldn't always prefer the same toys. Berta likes the fact that her son likes reading more than watching TV. I like the fact that Fabio likes what I like. Ronny prefers not to have such a strong preference for chocolate. The boyfriend says, "I'd like to do whatever you want to do", whereas his estranged lover thinks, "Oh, but right now I prefer doing whatever it is that you hate [...]."

As these examples show, second (and even higher) order wants, preferences and desires are a normal part of our life, although sometimes they get us into trouble. The so-called *paradox of desire* (I desire not to have

Received June 15, 2025. *Accepted* January 29, 2026. *Published online* March 1, 2026

the desires I have) is a case in point. But there are others. There is a particularly damaging combination of higher order likes and preferences in the context of games, which results in players having *ungrounded* or *paradoxical payoffs*.

This paper proposes the use of a dynamic logic of preferences to deepen our understanding of ungrounded payoffs and the need to seek for grounded interaction. The primary aim of this work is representational: to articulate a formal framework that captures the core intuitions of the problem. Developing a robust and insightful formal apparatus is a necessary preliminary step before pursuing more substantial formal results, which are left for future research.

I originally introduced the notion of ungrounded payoffs in an earlier paper (see Cresto, 2022). In what follows I will briefly revisit some key aspects of that proposal to clarify the approach I am pursuing here.

1. Ungrounded payoffs

Suppose Row and Column are invited to a party and are told to bring cheese and wine, respectively. Row can choose among a variety of types of cheese, and Column among types of wine. Row likes any particular combination of cheese and wine exactly to the extent that Column does (he thinks Column is more knowledgeable on these matters). Actually, in case Row likes a particular combination, what Row likes about it is that Column likes it. In addition, suppose now Row comes to believe that, according to Column, it is *Row's judgment* the one we should trust. In other words, Row also takes Column's preferences to be the result of Column's own conjectures on Row's preferences. Suppose that the situation is symmetrical for Column. Then, Row's and Column's preferences are not only interdependent, but ungrounded.

It is an empirical question whether such interdependence arises or not.[1] If it does arise, it threatens the presupposition that agents have an antecedent, well-defined von Neumann-Morgenstern utility function. As a result, it is not obvious how to fix a matrix to represent the interaction of the two agents in game theoretic terms.

Not all interactions among agents with interdependent preferences are equally problematic. The claim that, in some scenarios, agents with

[1] Kripke made a similar point long ago regarding the possible interdependence of truth values (see Kripke, 1975).

interdependent utility functions might fail to have a matrix requires some explanation. Possibly the clearest examples (but by no means the only ones) involve what we may call "perfect altruists" — or "perfect lovers", as I will call them from now on — and "perfect haters".[2] To simplify, suppose perfect lovers or haters only have either maximally preferred or maximally dispreferred outcomes — say, 1 or 0, if we adopt a normalized utility scale. A perfect lover obtains maximum payoff from an outcome if and only if she believes her partner also obtains maximum payoff, whereas perfect haters obtain 1 if and only if they believe their partners obtain 0.[3]

If perfect lovers or perfect haters believe they are facing another perfect lover or hater, the resulting situation might be impossible to be recast as a game. Note, first, that a couple of perfect lovers could have any matrix in which cells are pairs of payoffs of the form $(1, 1)$ or $(0, 0)$, whereas a couple of perfect haters could have any matrix with pairs of payoffs $(1, 0)$ or $(0, 1)$. Thus, for example, if Row and Column have two strategies each, the resulting 2x2 matrix for two perfect altruists could be any of the 16 possible matrices we can build out of $(1, 1)$- or $(0, 0)$-cells, whereas the 2x2 matrix for two perfect haters could be any of the 16 possible matrices we can build out of $(1, 0)$- or $(0, 1)$-cells. Although it might be possible for the agents to agree on which matrix is the one that reflects their interaction, they might also fail to do so (more on this later). On the other hand, there is no pair of payoffs that can reflect an interaction between a perfect altruist and a perfect hater, so no possible matrix can do the trick (at least when agents have only maximally preferred or dispreferred outcomes). In the last case payoffs are not only ungrounded, but paradoxical.

[2] For useful distinctions between psychological altruism (or spite) and biological or behavioral altruism (see Kitcher, 1993, 2010).

[3] A standard way of proceeding within the so called "other-regarding preferences" field in economics is to distinguish between the objective payoff matrix of an agent (for example, in terms of monetary units), and the subjective matrix of the altruist (see Bicchieri, 2006; Fehr and Schmidt, 1999; Rabin, 1993). However, in our examples agents do not have this possibility. The interdependence of preferences is radical, in the sense that there is no prior, objective matrix with objective payoffs the agents could rely on; we are dealing with von Neumann-Morgenstern utilities all along. An interesting early discussion of this phenomenon can be found in (Estlund, 1990), although he does not offer a formal reconstruction. His analysis arises in a very different context — that of mutual benevolence and Joseph Butler's theory of happiness.

		Column	
		Wine I	Wine II
Row	Cheese I	x, x	y, y
	Cheese II	z, z	u, u

Figure 1. The cheese and wine example, for the simplified case in which players only consider two brands of cheese and two brands of wine. We don't know what x, y, z or u are. Actually, there is no way we can know: the matrix is objectively underdetermined.

We may be tempted to model the aforementioned interactions of perfect lovers and haters as games with incomplete information. We may reason as follows: as Row's preferences depends on Column's, Row may seek to assign probabilities to Column's payoffs. Say, Row may think Column obtains a maximally preferred state of affairs under strategy profile a with probability x, and a maximally dispreferred state of affairs with probability $1 - x$. Then the expected payoff for Row under a is x, for $x \in [0, 1]$ (mutatis mutandis for Column). Unfortunately, this is not a viable route. The problem is that, as we have pointed out, in our example Row takes Column's preferences to be the result of Column's own conjectures on Row's preferences, which are not fixed, and vice-versa. This assumption is crucial, regardless of whether each agent can or cannot truthfully assume that their fellow player is or is not of her same "type" (a lover or a hater). Column's (Row's) preferences are just not well defined prior to Row's (Column's) conjectures. In other words, preferences are not only unknown, but ontologically undetermined.

In previous work I argued that, in some cases (but not in all cases) ontologically undetermined preferences can be fixed if the agents manage to implicitly coordinate their beliefs about the game they might be playing (Cresto, 2022). In particular, pairs of perfect altruists or haters (what we may call 'homogeneous' couples) can get involved in a second-order game in which they seek to coordinate the matrix of their first order interaction. In such cases, probability assignments to each other's preferences have a constructive character: rather than guides for players to discover the game they are playing, they are understood as tools to build the relevant matrix in the first place, out of what we will call "an underspecified game." In other words, there are interactions for which von Neumann-Morgenstern utilities do not pre-exist, but can be created, if agents achieve an equilibrium in pure strategies in a second-order coordination game. If coordination fails to occur, the first order interaction turns out not to be a game, in the technical sense of this term.[4]

This approach requires the assumption that players agree on the set of matrices available to them to play a first order game. Such an agreement can only occur if players are of the same type, so to speak; this is impossible for mixed couples of lovers and haters. In this paper I want to generalize my analysis of interdependent utility functions, by relaxing this assumption. As we will no longer assume that players agree on the set of possible matrices, it is no longer obvious that players will manage to play a second order game of the sort described above. How can we think of the process by which players try to fix a matrix, then, in the context of ungrounded payoffs? I will seek to model this process with the aid of a dynamic logic that can describe doxastic and preference attitudes of the players.

2. The proposal

In a nutshell, my main goal in this paper is to reconstruct the logical process by which players caught up in ungroundedness look for a suitable matrix, without presupposing that agents agree beforehand on which matrices are available to them. To accomplish this task I will rely on a system of dynamic logic with enough resources to express the beliefs and preferences of the agents. I will follow some of the ideas pioneered

[4] It may be possible to translate some aspects of this proposal into the framework of epistemic game theory, albeit with important caveats. Brandenburger and Keisler (2006) prove the impossibility of certain mutual belief configurations between agents; one could consider generalizations of their logic to address the indeterminacy of payoff matrices. Gul and Pesendorfer (2016) develop a model in which agents' preferences are sensitive to the perceived dispositions of others (e.g., altruism or spite), yet the interdependence of preferences remains limited and does not involve mutual fixation. To do the trick, one might begin by defining types for each player, where each player is associated with a belief hierarchy, as is standard in epistemic game theory. However, unlike traditional models where types are primarily defined by probabilistic beliefs about strategies and types of others, the present approach requires types to encode specific payoff functions as well. Please refer to (Cresto, 2022) for further details. A key difference between the modeling I favor and standard epistemic game-theoretic elaborations lies in the use of a second-order game structure, which permits breakdowns or failures in interaction — a feature I regard as essential. For general overviews of epistemic game theory, see (Bonanno, 2015; Dekel and Siniscalchi, 2015; Perea, 2012).

by Liu (2008), van Benthem, Otterloo and Roy (2006), van Benthem and Liu (2007, 2016), and Liu, Seligman and Girard (2014), although in this opportunity I will not deal with preference upgrades. Formally, the system I favor departs from older proposals of dynamic doxastic logic only in minor respects; however, the application is novel, and will help us identify a number of insights linking higher order attitudes and games.

Very sketchily, the project goes as follows. I will start by defining a multi-agent Kripke model for a possible 2-person game. I will talk irrespectively of 'players' or 'agents'. The possible worlds in the model discriminate among different strategy profiles and pairs of payoffs. I will also define different relations between worlds, which aim to capture the behavior of a belief and a preference operator of the language; the preference relation of each player will be determined 'endogenously', so to speak, on the basis of particular features of the worlds (namely, the payoffs), in a way to be explained. In addition, players will be able to upgrade their beliefs as a result of various inputs. Intuitively, players can change their beliefs on which world(s) may be the actual world, that is, which world(s) are good candidates to be the world in which the game is being played (in case there is such a thing). Players succeed in having a matrix if, as a result of successive upgrades, and for at least one strategy profile, they both end up believing that they are in the same world. This may fail to happen, either because of lack of coordination, or because of deeper impossibility. In such cases the verdict is that the agents are not playing a game in any possible world. Part of our task in the following sections will be to investigate conditions under which a sequence of upgrades comes to an end, and conditions under which a loop is guaranteed to occur.

Let me emphasize that the project just described does not attempt to reconstruct the reasoning of the players *within a game*; in particular, it does not attempt to reconstruct the logical process that leads players to reach a Nash equilibrium. In this sense, the present use to dynamic logic is very different from other proposals found in the literature that relate dynamic logic and games. Recall that our agents are not yet playing a game — and, indeed, they might end up discovering that playing a game is an impossible task under the circumstances. More generally, the proposal can be framed as a sort of *mental pre-play of a game*. In general, this mental stage—what happens in one's mind before the actual game unfolds — can be a powerful tool for reasoning strategically.[5] Modeling a

[5] Thanks to an anonymous reviewer for suggesting this framing for my proposal.

mental pre-play can be interesting in its own right, beyond its usefulness in addressing ungroundedness, and may lend itself to a variety of future applications.

In what follows I will continue to focus on the fate of perfect lovers and haters, to make examples more concrete and easier to grasp. We should bear in mind, however, that ungroundedness is not an exclusive phenomenon of altruistic or spiteful attitudes, as the Cheese and Wine example already taught us.

3. The basic framework

In this section I will present the basic logic framework for a static setting; we will incorporate dynamic elements in further sections. Recall that our goal is to be able to talk about players involved in interactions in which the matrix is radically underdetermined, so we need some preliminaries first. To this effect, consider the concept of an *underspecified game*:

Definition 1. An *underspecified game* $G^U = \langle I, S, \pi, r\rangle$ is a tuple where:

- $I = \{1, \ldots, n\}$ is the set of players.
- $S = \{S_1, \ldots, S_n\}$ contains sets of strategies; for every $i \in I, S_i = \{s^i_1, \ldots, s^i_m\}$ is a set of strategies for i, where each strategy s^i_k in S_i is a fully specified course of action for player i. A *strategy profile* is a tuple of strategies $\langle s^i, \ldots, s^n\rangle$.
- $\pi = \langle\{\pi^\gamma_1\}, \ldots\{\pi^\gamma_n\}\rangle$ is a vector of *sets* of von Neumann-Morgenstern utility functions $\{\pi^\gamma_i\}_{\gamma=1,\ldots}$, for each $i \in I$, where each π^γ_i ranges from strategy profiles to real numbers.
- For every strategy profile a, $\langle\pi^\delta_1(a), \ldots, \pi^\lambda_n(a)\rangle$ is a tuple of payoffs, where, for every $i, \pi^\delta_i \in \{\pi^\gamma_i\}$.
- r is a (possibly empty) set of restrictions r_i, for $i \in I$, where each r_i provides constraints on tuples of payoffs $\langle\pi^\delta_1(a), \ldots, \pi^\lambda_n(a)\rangle$, for each strategy profile a.[6]

I will often use lowercase $a, b, \ldots$ to refer to strategy profiles. Note that π determines a space of possible matrices, which may be made smaller by the set of restrictions. Restrictions $r_1, \ldots, r_n$ capture the way in which each player believes that her own utility function interacts

[6] More formally, r is a vector $\langle r_{i,a}\rangle_{i\in I, a\in S} \subseteq \mathbb{R}^I$, meaning that for all $i \in I$ and $a \in S$, the vector of payoffs $\langle\pi^\delta_1(a), \ldots, \pi^\lambda_n(a)\rangle \in \mathbb{R}^I$ is allowed iff it is in $r_{i,a}$.

with the functions of others, and hence they jointly determine a set of *admissible* matrices for the players, which is a (not necessarily proper) subset of all possible matrices. Restrictions need not be enough to pin down a unique matrix; on the other hand, if restrictions cannot be jointly satisfied, the set of admissible matrices of G^U will be empty.

For the rest of the paper we will focus on a particular family of underspecified games[7] g^U, such that: (i) $I = \{1, 2\}$; (ii) there are two possible strategies for each player: $A1$ and $B1$ for player 1, and $A2$ and $B2$ for player 2; let 'a', 'b', 'c' and 'd' refer to the four resulting strategy profiles, where $a = \langle A1, A2\rangle$, $b = \langle B1, A2\rangle$, $c = \langle A1, B2\rangle$ and $d = \langle B1, B2\rangle$; (iii) for any player i and any strategy profile $k, \pi_i(k) = \{0, 1\}$; and finally, (iv) for any player i, restriction r_i can be of one of two possible types, in the following way:

- r^L: for every strategy profile k, $\pi_i(k) = \pi_j(k)$;
- r^H: for every strategy profile k, $\pi_i(k) = 1 - \pi_j(k)$ (for $i \neq j$).

Intuitively, a player has restriction r^L if she is a perfect lover, and restriction r^H if she is a perfect hater. Note that 'homogeneous' couples of perfect lovers or perfect haters restrict the set of all possible matrices (256 matrices, for payoffs in {0, 1}) to 16, whereas the constraints imposed by a mixed couple cannot be jointly satisfied for $\pi_i(k) = \{0, 1\}$, and hence in this case the space of admissible matrices is empty.[8]

Now we introduce a language L to describe the agents in g^U, in the usual way:

$$\phi ::= \ \bot \mid p \mid \neg\phi \mid \phi \wedge \psi \mid B_i\phi \mid \langle \mathrm{Pref}\rangle_i\phi \mid \langle \mathrm{DisPref}\rangle_i\phi.$$

'$B_i\phi$' stands for 'agent i believes that ϕ', for $i = 1, 2$. Here I will only deal with plain beliefs that do not admit of degrees, and will not seek to analyse how belief in this basic sense relates to credences or other graded accounts. Note that agents have a very elementary preference structure, as outcomes are maximally preferred or maximally dispreferred. This simplifies the intended reading of '$\langle \mathrm{Pref}\rangle_i\phi$', which can be just rendered

[7] I use lowercase 'g^U' to refer to underspecified games G^U in which the following conditions hold.

[8] These restrictions do not capture all we need to capture about perfect lovers and haters. A perfect lover [hater] has payoff 1 because s/he *believes* her partner has payoff 1 [0]. The doxastic component does not make any difference at the time of describing the set of admissible matrices, but of course it will be relevant at the time of understanding the dynamic aspects of the model, as we will see in a moment.

as 'agent i likes ϕ', whereas '$\langle \text{DisPref} \rangle_i \phi$' simply means 'agent i dislikes ϕ'. Richer accounts of preference and belief can be added, of course, but at this point the complications will not pay off, and I fear they might obscure the way the framework is meant to help us achieve our goals.[9]

Consider now the following model for L:

$$\mathcal{M} = \langle W, (R_i)_{i \in I}, (\preceq_i)_{i \in I}, V \rangle,$$

where W is a set of worlds. Worlds should be able to discriminate strategy profiles and pairs of payoffs; there should be at least as many worlds as strategy profiles, of course, but two worlds can capture the same strategy profile and pair of payoffs and still be distinct.[10] Worlds can be conventionally represented by an indexed pair (with possible indices $i = 1, 2, \ldots$) consisting of a profile and a pair of payoffs, as in, say, $w = (a, (1, 1))_5 \in W$; $z = (b, (1, 0))_8 \in W$, etc.

Each R_i is a serial, transitive and Euclidean relation between worlds. This will ensure that the B operator behaves as in the KD45 modal system. This simplifies a bit the plausibility models that are normally offered by dynamic doxastic logic; intuitively, in our setting the accessible worlds correspond to the most plausible ones.

On the other hand, $\preceq_i$ is the relation between worlds induced by π_i, in the following manner. Let x be i's payoff for strategy profile k in world s, and let y be i's payoff for strategy profile l in world t (where 'k' and 'l' may, but need not, be distinct). Then $s \preceq_i t$ if and only if $x \leq y$; from this, a strict preference relation can be defined in the obvious way. We read '$s \preceq_i t$' as 't is at least as good for agent i as s'. As '$\pi(k)$' can adopt only two possible values, for any k, it induces a very simple preference relation where players have preferred and dis-preferred worlds (or worlds they like and worlds they do not like): if someone has payoff 0 in world s and payoff 1 in world t, then she will think that t is strictly better than s.

Finally, V is a valuation function for atomic formulas. The valuation of sentences of L then proceeds inductively in the usual way:

[9] Note that, as is usually the case in logics applied to games, the language need not have resources to express all metalinguistic truths about a game (or, in this case, about an underspecified game).

[10] It is standard practice that worlds need not *identify* with strategy profiles. See for example (Stalnaker, 1994).

Definition 2. Given $\mathcal{M} = \langle W, (R_i)_{i \in I}, (\preceq_i)_{i \in I}, V \rangle$, and a world $w \in W$, we define $\mathcal{M}, w \models \phi$ by induction on ϕ:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$.
- $\mathcal{M}, w \models \neg\phi$ iff not $\mathcal{M}, w \models \phi$.
- $\mathcal{M}, w \models \phi \wedge \psi$ iff $\mathcal{M}, w \models \phi$ and $\mathcal{M}, w \models \psi$.
- $\mathcal{M}, w \models B_i\phi$ iff for all y: if wR_iy, then $\mathcal{M}, y \models \phi$.
- $\mathcal{M}, w \models \langle \text{Pref} \rangle_i \phi$ iff for some y: $w \preceq_i y$ and $\mathcal{M}, y \models \phi$.
- $\mathcal{M}, w \models \langle \text{DisPref} \rangle_i \phi$ iff for some y: $y \preceq_i w$ and $\mathcal{M}, y \models \phi$.

In the model just defined, an agent is of a certain *type* if she can access certain worlds and not others, i.e., if she deems possible that the actual world is a world in which certain preferences hold, and not others. So a type can be partially described by a collection of statements that mention beliefs and preferences. In what follows we will stipulate that players of g^U can be of one of two types. Let 'p' be true exactly in the worlds in which strategy profile a holds.

Perfect lovers. Perfect lovers can only access (0,0)- or (1,1)-worlds. In addition, for any a-world w, i is a Perfect lover if and only if

- $\mathcal{M}, w \models B_i \langle \text{Pref} \rangle_i p$ iff $\mathcal{M}, w \models B_i \langle \text{Pref} \rangle_j p$ iff $\mathcal{M}, w \models B_i B_j \langle \text{Pref} \rangle_j p$,

from which it is easy to see that,

- if $\mathcal{M}, w \models B_i \langle \text{Pref} \rangle_i p$, then $\mathcal{M}, w \models \langle \text{Pref} \rangle_i \langle \text{Pref} \rangle_j p$.

Perfect haters. Perfect haters, on the other hand, can only access (1,0)- or (0,1)-worlds. In addition, for any a-world w, i is a Perfect hater if and only if

- $\mathcal{M}, w \models B_i \langle \text{Pref} \rangle_i p$ iff $\mathcal{M}, w \models B_i \langle \text{DisPref} \rangle_j p$ iff $\mathcal{M}, w \models B_i B_j \langle \text{DisPref} \rangle_j p$;
- $\mathcal{M}, w \models B_i \langle \text{DisPref} \rangle_i p$ iff $\mathcal{M}, w \models B_i \langle \text{Pref} \rangle_j p$ iff $\mathcal{M}, w \models B_i B_j \langle \text{Pref} \rangle_j p$,

which entails that

- if $\mathcal{M}, w \models B_i \langle \text{Pref} \rangle_i p$, then $\mathcal{M}, w \models \langle \text{Pref} \rangle_i \langle \text{DisPref} \rangle_j p$;
- if $\mathcal{M}, w \models B_i \langle \text{DisPref} \rangle_i p$, then $\mathcal{M}, w \models \langle \text{DisPref} \rangle_i \langle \text{Pref} \rangle_j p$.[11]

[11] Note that types generate infinite hierarchies of beliefs and preferences. For a more classical approach to handling preference hierarchies in games, see (Di Tillio, 2008).

In the light of all this, note that describing the preference relation of an agent requires paying attention to two very different things:

- Her first order preferences on strategies;
- Her higher order preferences, related to types: an agent likes or dislikes worlds depending on whether those worlds are liked or disliked by her partner.

More sophisticated settings might allow us to define other types as well, or may allow for the possibility that agents do not have a defined type at all; we might also want to allow agents to change their types depending on various factors. Here, however, I will stick to the simpler scenario, and assume that (i) types are stable; and that (ii) it is common knowledge among the players that each player is either a Perfect Lover or a Perfect Hater.

Types within dynamic epistemic logic have already been explored by Liu (2009), in the context of a research on epistemic and doxastic updates driven by others' opinions, observation, and memory, though not preferences. In particular, she introduces predictates to account for *liars* and *truth-tellers* (see section 6, "Interaction between different agents"), as well as for *introspective* and *non-introspective* agents; her framework also accommodates uncertainty regarding agent types. Unlike her proposal, in this opportunity I will not seek to express the types of players in the language; I will leave this possibility for future work.

As we have seen, types provide constraints for accessibility relations (i.e., constraints for the underlying doxastic structure of the model). But notice that, by doing so, types also encode revision policies, in the following sense. Suppose that as a result of a doxastic change by j, there is now a world in which player i ends up believing both that she likes p-worlds, and that j does not like them; in other words, suppose a model is upgraded in such a way that now $\mathcal{M}, w \models B_i\langle \text{Pref}\rangle_i p$ and also $\mathcal{M}, w \models \neg B_i\langle \text{Pref}\rangle_j p$. If i is a Perfect Lover, this situation creates a tension, which triggers endogenously a new revision. This gives us some hints on how the dynamics of the model should go.

Before introducing the dynamic setting, there is a further point I would like to stress. As usual when modeling games, which world is actual depends in part on what the two agents choose to do. In our current predicament, however, it is not obvious whether some of the worlds in W is indeed actual. And by this I do not mean to say that agents may *ignore* which world is actual. Recall that the actual world

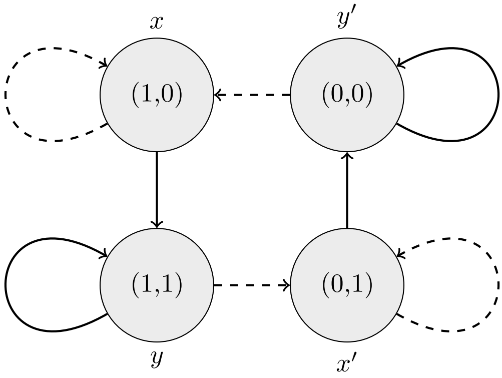

Figure 2. A partial rendering of $\mathcal{M}$, for strategy profile a

should be the one in which the agents are playing a game. But that might turn out to be an impossible world, and hence not in the algebra in the first place (at any rate: not in *this* algebra).

By way of illustration, consider the following partial rendering of a possible model $\mathcal{M}$, for strategy profile a (Figure 2). The diagram is only partial, as I have only included a-worlds in it; this will suffice for present purposes. Here solid arrows represent the doxastic accessibility relation of player 1, while dashed arrows represent that of player 2. Notice that player 1 is a Perfect Lover and player 2 a Perfect Hater: player 1 can only access worlds y and y', while player 2 can only access x and x'. In world y, player 1 believes that both 1 and 2 like world y; however, at world y player 1 also believes that player 2 wrongly believes to be in a world player 1 dislikes (i.e., world x'). *Mutatis mutandis*, this situation repeats in every world: for all worlds w, player i believes that player j is mistaken about the world they are both in.

4. The dynamic setting

We are now ready to incorporate dynamic aspects to our model. In principle, a doxastic state can change in response to so-called soft information about:

- The strategy the other player is attempting to play; or
- The payoff of the other player under certain strategy profile, which is believed to be instantiated (i.e., the *preferences* of the other player).

Following usual terminology, I will refer to such changes as 'doxastic upgrades' (see, e.g., van Benthem and Smets, 2015). Doxastic upgrades can remove some accessibility links and create others. As we are dealing exclusively with plain beliefs (i.e. not probabilistic), doxastic upgrades in our setting can only take the form of 'radical upgrades', where a radical upgrade moves all ϕ-worlds before all the not-ϕ-worlds.

The formal machinery I present below could be supplemented with tools to allow for preference shifts, but we will not need them in the present context. In any case, as I pointed out before, note that belief changes can indirectly affect first order preferences on strategies, in the sense that, given a particular strategy, agents can end up accessing only worlds where that strategy is preferred, or dis-preferred by her partner.

The general picture is again the standard one. There are triggers that, in each case, change some of the doxastic relations of at least one of the players, yielding a new, revised model:

$$\mathcal{M}_1 \xrightarrow{\tau_1} \mathcal{M}_2 \xrightarrow{\tau_2} \mathcal{M}_3 \xrightarrow{\tau_3} \ldots \xrightarrow{\tau_n} \mathcal{M}_{n+1} \ldots$$

As is normally the case when dealing with the application of dynamic logic to games, the resulting sequence of models is not meant to be understood chronologically, but it is rather a logical reconstruction of the reasoning of the players.[12] In our current setting, it is a reconstruction of the logical steps taken by agents who are searching for a matrix.

We present now the full formal machinery. Our language for the dynamic setting incorporates action expressions for announcements, in the usual way:

$$\phi := \ \bot \mid p \mid \neg\phi \mid \phi \wedge \psi \mid B_i\phi \mid \langle \text{Pref} \rangle_i \phi \mid \langle \text{DisPref} \rangle_i \mid U\phi \mid [A]\phi,$$
$$A := \ \Uparrow\phi.$$

Triggers for doxastic upgrades are events in which ϕ seems to be the case, with some degree of uncertainty, which is normally written as '$\Uparrow\phi$':

Definition 3. A radical upgrade $\Uparrow\phi$ changes the current accessibility relation R_i between worlds in $\mathcal{M}, w$, to create a new model $\mathcal{M}_{\Uparrow\phi}, w$

[12] A very clear rendering of this methodology can be found for example in (Pacuit and Roy, 2016, p. 305): "[T]he idea is to represent the process of *rational deliberation* that takes the players from the *ex ante* stage to the *ex interim* stage of decision making. Thus, the "informational exchanges" are the result of the players' practical reasoning about what they should do, given their current beliefs."

where all ϕ-worlds in $\mathcal{M}, w$ become accessible from w and all $\neg\phi$-worlds become inaccessible.

Thus, $\mathcal{M}, w \models [\Uparrow\phi]\psi$ iff $\mathcal{M}_{\Uparrow\phi}, w \models \psi$.

As I have already anticipated, in many interesting cases doxastic triggers will be generated endogenously, and will allow for what we may call *upgrade cascades*. Given that for each player preferences on worlds are determined by the preferences of the other player, by changing her beliefs about the preferences of player j, player i may at the same time change her beliefs about her own preferences. Thus, a shift in the accessibility relations of player j (say, in model $\mathcal{M}_n$), when noticed by i, may have an immediate effect on the accessibility relations of player i (in model $\mathcal{M}_{n+1}$). The upshot is that an initial upgrade can create a chain of doxastic shifts.

Very informally, an upgrade cascade works as follows. Suppose that at the very beginning players i and j are in complete suspension of judgment. This does not mean, however, that they can access every possible world, because the set of worlds each of them 'sees' is restricted by his or her type. Now suppose something player i does or says leads j to understand that i will perform strategy $A1$, and that i actually likes only $A1$-worlds. Let us call this 'the initial background for upgrade'. The initial background can consist of (more or less subtle) verbal or behavioral cues by i, but this is not necessary; it could also involve other events in the world (including the reports of other agents) that give j hints on the likes and dislikes of player i, or even j's recollection of other past events, which are now re-signified. Be that as it may, the background catalyzes in a particular piece of soft information, which we will call 'the initial trigger' or 'initial input'; the initial trigger is an event that starts an upgrade process. In our example, the initial input is the soft announcement that i likes $A1$-worlds.[13]

Suppose that j is a perfect lover. Then j likes the same worlds i likes. Recall that we abbreviate $a = \langle A1, A2\rangle$; $b = \langle B1, A2\rangle$; $c = \langle A1, B2\rangle$; and $d = \langle B1, B2\rangle$. As a result of an initial input, suppose j can now only ac-

[13] An interesting antecedent to the present proposal can be found in (Seligman, Liu and Girard, 2011) and (Liang and Seligman, 2011), where explore how individual preferences can influence others within a community. They show how the "contagion" of preferences can give rise to cycles, which may eventually stabilize, although their analysis does not address the problem of ungroundedness. Liu, Seligman and Girard (2014) further investigate the flow of social information by relying on the framework of public announcements, albeit in the context of beliefs rather than preferences.

cess worlds characterized by $(a, (1, 1))$, $(b, (1, 1))$, $(c, (0, 0))$, or $(d, (0, 0))$. In other words, the initial input produces a shift in j's doxastic relation and hence determines a new, second model. But the new model makes i aware of this very fact. We may call this an 'endogenous' trigger. The new model automatically generates a new trigger endogenously, which can be captured by the soft announcement that j likes (only) a-worlds and b-worlds. This new trigger may prompt a change in player i's doxastic structure, depending on i's own type, and therefore it determines a third model. For example, if i is a perfect hater, (soft)-learning that j likes a-worlds may lead player i to dislike them, and to believe this much. This may go on and on, or eventually stop at some point.

The upgrade is 'soft' in the sense that players can get things wrong; they can misunderstand the relevant cues. After the initial external input, the sequence of endogenous doxastic changes is determined by the underlying structure of the model, until a new external input appears. Still, as the whole sequence is the consequence of an initial piece of uncertain information, all subsequent changes are also 'soft'.

Let us see the dynamic models in action. Consider, for starters, a pair of perfect lovers. To simplify the analysis we assume once again that agents already believe that they are living in some a-world (for some strategy profile a); again, let 'p' be true exclusively in all worlds in which strategy profile a holds. Upgrades on which strategy is believed to be implemented is quite standard and will not be revised here. As they are both perfect lovers, no player can live (or access) a (0,1)- or (1,0)- world. Figure 3 shows the initial setting, when both payers are in suspense regarding which a-world is the actual world. (As usual, the rectangle is meant to represent an equivalence relation, so within its limits all worlds access every other world.)

Now suppose that both agents receive the soft information that the other player actually *likes* a-worlds. The trigger prompts an upgrade, as a result of which we obtain the revised model shown in Figure 4. Interestingly, this model is 'stable', in the sense that both players found a world which they both believe to be the actual world.[14]

Consider now a different scenario. Suppose instead that they both

[14] Recall that 'stability' here means something very different from what is normally understood in the context of games, as the players are not yet playing a game — and in the end they might be unable to do it. In the last section I will discuss why it is relevant for the players to believe that they are both in the same world, conditional on believing a particular strategy profile holds.

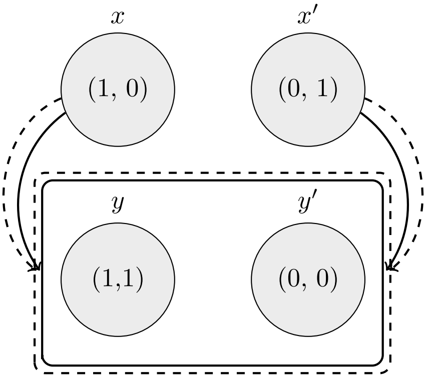

Figure 3. Perfect lovers – initial setting

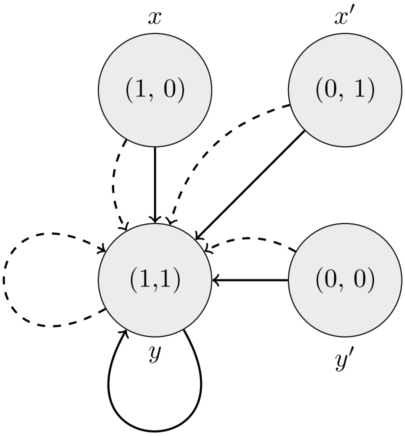

Figure 4. Perfect lovers – a stable revision trigger: $\Uparrow\langle\text{Pref}\rangle_1 p$, $\Uparrow\langle\text{Pref}\rangle_2 p$

misunderstand the cues; alternatively, it may happen that player 1 genuinely tries to signal that she likes a-worlds, whereas player 2 genuinely tries to signal that she dislikes them. As a first response to the trigger, the upgrade will lead players to the model represented in Figure 5. But at that stage they will both change their accessibility relations once again, because, being perfect lovers, they cannot (dis)like a strategy unless their partner (dis)likes it as well, which will bring us to Figure 6.

Unfortunately, as it is easy to see, the players cannot remain in Figure 6 either: the new model generates new triggers, as a result of

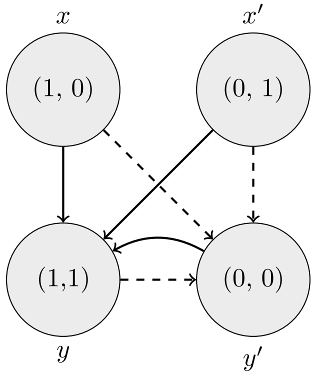

Figure 5. Perfect lovers – unstable revision (I) trigger: $\Uparrow\langle\text{Pref}\rangle_1 p$, $\Uparrow\langle\text{DisPref}\rangle_2 p$

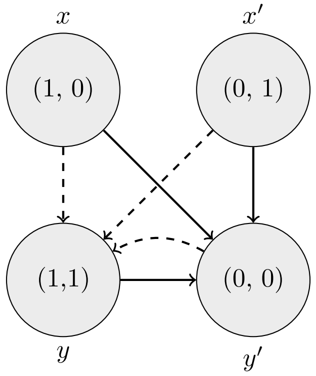

Figure 6. Perfect lovers−unstable revision (II) trigger: $\Uparrow\langle\text{DisPref}\rangle_1 p$, $\Uparrow\langle\text{Pref}\rangle_2 p$

which the two players will revert to their doxastic state from the previous model, and then they will go back and forth between the two. In other words, the players are caught up in a loop.

The analysis for pairs of perfect haters runs along similar lines. Things become more complex, but also more interesting, if we have a mixed couple in which a Perfect Hater faces a Perfect Lover. Consider an initial setting as shown in Figure 7. Here the solid-line player is a Perfect Lover, so she cannot access a (0,1)- or (1,0)-world, while the dashed-line player is a Perfect Hater, who cannot access (1,1)- or (0,0)-worlds. So far the setting is stable, because there is crucial lack of information: each player ignores the valuations of others. The solid-line

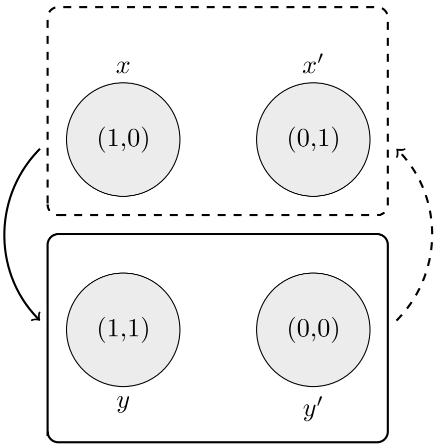

Figure 7. Perfect altruist meets a Perfect Hater: initial setting

player believes that the dashed-line player does not know whether the solid-line player prefers a-worlds, and vice versa. Players are in state of maximal ignorance within a-worlds. The stability is broken as soon as they receive externally more (soft) information about each other.

Suppose the initial trigger is the 'announcement' that player 2 likes a-worlds, i.e., $\Uparrow\langle\text{Pref}\rangle_2 p$. As a result of this, the solid-line player now changes her accessibility relations, and she now thinks she is in world $y = (a, (1, 1))$. But then the dashed-line player has the new input that the solid-line player likes a-worlds as well, i.e., $\Uparrow\langle\text{Pref}\rangle_1 p$ (Figure 8). Hence the dashed-line player 'moves' to world x, which provides a new trigger for the solid-line player, namely, that dashed-line player dislikes p (Figure 9). Solid-line player then 'moves' to world y', which provides a further trigger for the dashed-line player (Figure 10). As a result of this, the dashed-line player changes her doxastic accessibility relations once more, and now can only access world x' (Figure 11). At this point it is clear that the players will start the same sequence of upgrades all over again. As we can see, players will keep on upgrading their beliefs forever, and in each and every upgrade, they are bound to believe that the other player is confused about the actual world.

Say that a player is opinionated on ϕ in a model $\mathcal{M}$ if and only if all worlds in $\mathcal{M}$ are such that ϕ holds. It is easy to see that mixed couples of perfect lovers and haters will generate a loop of upgrades regardless of the details of the initial trigger, as long as such a trigger suffices to

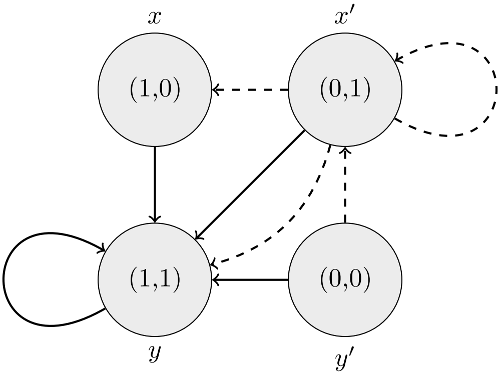

Figure 8. Perfect altruist meets a perfect hater trigger: $\Uparrow\langle \mathrm{Pref}\rangle_2 p$

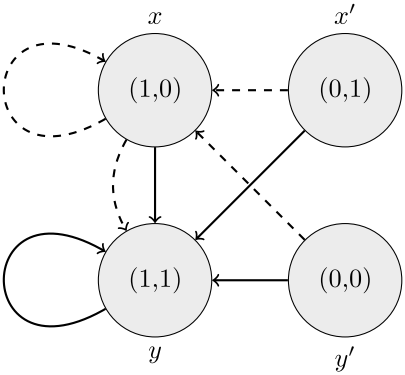

Figure 9. Perfect altruist meets a perfect hater trigger: $\Uparrow\langle \mathrm{Pref}\rangle_1 p$

make one of the players opinionated on whether she likes or dislikes a-worlds, for some strategy profile a. Moreover, this result holds even if players were still uncertain on whether they are, or are not, in an a-world to begin with. The loops we are considering here are 'strategy profile-dependent', in the sense that all it takes for them to arise is for a player to become convinced of her preferences over a particular strategy profile, *in case that particular strategy profile were to be instantiated.* (So the loop would remain the same even if players were still uncertain regarding which strategy profile will hold). We summarize these results in the following observation, which generalizes our previous example:

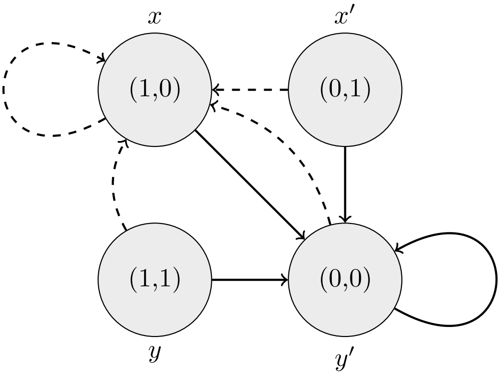

Figure 10. Perfect altruist meets a perfect hater trigger: $\Uparrow\langle\text{DisPref}\rangle_2 p$

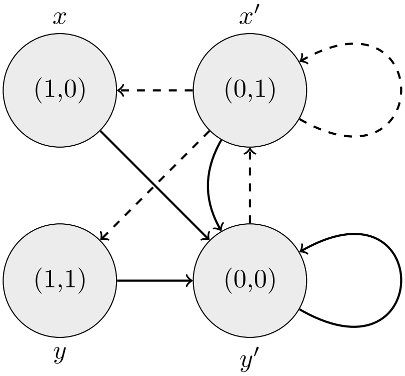

Figure 11. Perfect altruist meets a perfect hater trigger: $\Uparrow\langle\text{DisPref}\rangle_1 p$

Observation. Consider a multiagent Kripke model $\mathcal{M}$ as defined before, where the two agents are of fixed but opposite types. Then, for any sentence ψ:

- $[\Uparrow\psi]$ generates an infinite sequence of upgrades if there is at least a strategy profile a and a player j such that j is totally opinionated in $\mathcal{M}_{\Uparrow\psi}$ regarding her preferences for $p :=$ 'strategy profile a holds'.

5. Further results

So far we have shown that, for heterogeneous couples, as soon as they become opinionated regarding their preferences for some a-world, the two agents get trapped in an infinite sequence of upgrades; for each and every model, each agent is bound to believe that the other is confused about which a-world can be the actual world.

If we add further expressive resources to the language other results become possible as well, of course. For example, we may wonder if there can be common belief in the fact that it is impossible for the two agents to play a game, or, equivalently, if there can be common belief in the fact that there is no upgrade which will enable them to believe that they are at the same world.

To settle this question we should somehow manage to inject some information about structural features of the modeling framework into the object language. I will show that we can do that without substantive additions to the previous machinery. To this effect, consider the following definitions:

Let **g** be the proposition that the agents can fix a matrix and play a game. What **g** actually says is that there is some strategy profile a, and some world w within some upgraded model $\mathcal{M}_n$ (out of an initial model $\mathcal{M}_1$), such that either (i) both players believe that they like a in w, and they both believe that the other player likes it as well; or (ii) they both believe that they dislike a in w, and they both believe that the other player dislikes it as well; or (iii) agent i believes that she likes a in w and that j dislikes it, while agent j believes that she dislikes a in w and that i likes it; or (iv) agent i believes that she dislikes a in w and that j likes it, while agent j believes that she likes a in w and that i dislikes it.

On the other hand, let '$\mathbf{g_m}$' be the proposition that the agents can play a game *in model* $\mathcal{M}_m$, where $\mathcal{M}_m$ has been obtained from $\mathcal{M}_1$ as a result of a sequence of upgrades.

Note that '$\neg$**g**' says that it is *impossible* for the players to play a game, and not just that no game is being played in the particular model in which it is being evaluated: they will not find a matrix in any possible upgrade of $\mathcal{M}_1$. In other words, **g** is either true in every possible world, for every possible upgrade of $\mathcal{M}_1$, or false in every world, for every upgraded model. To say that there *can* be a matrix does not mean to say that the players have a matrix now, but that it is possible to find one, i.e., that there is a future path of revisions that eventually finds it.

On the other hand, $\mathbf{g}_m$ is either true in every world of $\mathcal{M}_m$, or false in every world of $\mathcal{M}_m$. In light of all this, the truth conditions of $\mathbf{g}$ and $\mathbf{g}_m$ are as follows:

Definition 4. Consider an initial model $\mathcal{M}_1 = \langle W, (R_i)_{i\in I}, (\preceq_i)_{i\in I}\rangle$, a world $w \in W$, and a model $\mathcal{M}_m$ that results from $\mathcal{M}_1$ after a sequence of m upgrades. Let $p :=$ 'strategy profile a holds', for some a. Then:

- $\mathcal{M}_m, w \models \mathbf{g}_m$ iff there is some world $v \in W$ such that
  - both believe they both like p in v: $\mathcal{M}_m, v \models B_i\langle\text{Pref}\rangle_i p \wedge B_j\langle\text{Pref}\rangle_j p \wedge B_iB_j\langle\text{Pref}\rangle_j p \wedge B_jB_i\langle\text{Pref}\rangle_i p$; or
  - both believe they both dislike p in v: $\mathcal{M}_m, v \models B_i\langle\text{DisPref}\rangle_i p \wedge B_j\langle\text{DisPref}\rangle_j p \wedge B_iB_j\langle\text{DisPref}\rangle_j p \wedge B_jB_i\langle\text{DisPref}\rangle_i p$; or
  - they both believe that i likes p and j dislikes p in v: $\mathcal{M}_m, v \models B_i\langle\text{Pref}\rangle_i p \wedge B_j\langle\text{DisPref}\rangle_j p \wedge B_iB_j\langle\text{DisPref}\rangle_j p \wedge B_jB_i\langle\text{Pref}\rangle_i p$; or
  - they both believe that j likes p and i dislikes p in v: $\mathcal{M}_m, v \models B_i\langle\text{DisPref}\rangle_i p \wedge B_j\langle\text{Pref}\rangle_j p \wedge B_iB_j\langle\text{Pref}\rangle_j p \wedge B_jB_i\langle\text{DisPref}\rangle_i p$.
- $\mathcal{M}_m, w \models \mathbf{g}$ iff there is some model $\mathcal{M}_n$ and some sequence of upgrades from $\mathcal{M}_1$ to $\mathcal{M}_n$, such that $\mathcal{M}_n, w \models \mathbf{g_n}$.

In other words, $\mathbf{g}$ is true in any upgraded model iff $\mathbf{g_n}$ is true, for some n. Note that $\mathbf{g}$ and $\mathbf{g_n}$ are just abbreviations that rely on previous elements of L; as a result, we are mimicking the effects of a hybrid logic machinery without actually enlarging our basic resources. The following observation is also straightforward:

Theorem 1. *The proposition* $\mathbf{g}$ *is necessarily true for homogeneous couples, and necessarily false for heterogeneous couples.*

Note that the players need not get involved in an infinite upgrade cascade in order to be unable to play a game; in Figure 2 we showed a case of a stable situation for an heterogeneous couple. In any case, the truth conditions for $\mathbf{g_n}$ are never met.

How can agents come to believe, or disbelieve, that $\mathbf{g}$ is the case? To capture this idea we define the concept of *belief-in-the-limit*, or '$B_{i,\rightarrow}$'.

Definition 5. For any model $\mathcal{M}_m$, any $w \in W$, any player i and any formula ϕ:

- $\mathcal{M}_m, w \models B_{i,\rightarrow}\phi$ iff for any n such that there is a path of possible upgrades from $\mathcal{M}_m$ to $\mathcal{M}_n$, $\mathcal{M}_n, w \models B_i\phi$.

In other words, a player i believes-in-the limit that ϕ, in a given model $\mathcal{M}_m$, if she believes that ϕ in every possible model that can be reached from $\mathcal{M}_m$ through some sequence of upgrades. Thus, belief-in-the-limit captures the idea that beliefs sometimes remain stable for any upgraded model as further away from the initial model as we want, for an n that maybe as large as we want.[15] Of course, logical truths are always believed-in-the-limit. But so are necessary truths that capture the structure of the interaction.

In what follows we define *common belief-in-the-limit* by analogy with the standard definition of common belief:

Definition 6. Mutual belief-in-the-limit:

- $\mathcal{M}_n, w \models E_{i,j,\to}\phi$ iff $\mathcal{M}_n, w \models B_{i,\to}\phi$ for all $i \in I$.

  Iteration of mutual belief-in-the-limit:
- $E^1_{i,j,\to}\phi := E_{i,j,\to}\phi$, whereas $E^{k+1}_{i,j,\to}\phi = E_{i,j,\to}E^k_{i,j,\to}\phi$

  Common belief-in-the-limit:
- $\mathcal{M}_n, w \models CB_{i,j,\to}\phi$ iff $\mathcal{M}_n, w \models E^k_{i,j,\to}\phi$, for $k = 1, 2, \ldots$.

With the aid of these tools, we can show that:

Theorem 2. *Let $\mathcal{M}_1$ be as before. If the agents in I are of different types, then, for every possible upgrade of $\mathcal{M}_1$, they have common belief-in-the-limit that they cannot play a game.*

In other words, 'heterogeneous' pairs of agents can anticipate instability — they can anticipate that they will be caught up in a paradox. The proof is again straightforward:

Proof. First we show that each player has a belief-in-the-limit in $\neg\mathbf{g}$.

Without loss of generality, we evaluate $B_{i,\to}\neg\mathbf{g}$ in model $\mathcal{M}_1$. By definition, $\mathcal{M}_1, w \models B_{i,\to}\neg g$ is the case iff for any n such that there is a path of possible upgrades that goes from $\mathcal{M}_1$ to $\mathcal{M}_n$, $\mathcal{M}_n, w \models B_i\neg g$. But, by Theorem 1, $\neg\mathbf{g}$ is true in every world of $\mathcal{M}_n$, and hence i believes it in $\mathcal{M}_n$. Hence $\mathcal{M}_1, w \models B_{i,\to}\neg\mathbf{g}$. The case for j is analogous. As both $B_{i,\to}\neg\mathbf{g}$ and $B_{j,\to}\neg g$ hold in every world of every upgraded model, the

[15] In a more sophisticated setting we could also allow that the B-operator be indexed to models, and that players can make reference to them within a given model. This would open up the possibility to express more complex propositions; for example, players may be able to recollect their past beliefs or predict future ones. This possibility is interesting in itself, but I will not explore it here.

two agents believe it, they believe that they believe it, and so forth. Hence $CB_{i,j,\to}\neg\mathbf{g}$ holds as well. $\dashv$

On the other hand, recall that in order to play a game, not only should R_i and R_j have a particular structure, but also, the upgrades need to be felicitous; as we have seen in Figures 5 and 6, this may fail to be the case even for homogeneous couples.

Corollary. *Let $\mathcal{M}_1$ be as before. If the agents in I are of the same type, then, in every possible upgrade of $\mathcal{M}_1$ they have common belief-in-the-limit that they can play a game. However, for certain initial models and triggers, unlucky couples believe-in-the-limit that* $\mathbf{g}$*, but believe that $\neg\mathbf{g}_n$ in every particular upgraded model $\mathcal{M}_n$.*

This may sound slightly paradoxical, but it is not. The reason this is consistent is that $\mathbf{g}$ is actually stronger than the conjunction of the $\mathbf{g}_n$'s that can be accessed by unlucky couples — the latter depends on the specific details of the specific path of upgrades that actually occurred.

If we add predicates for agent types to the language, we can further explore whether an agent can come to believe that her partner is of a certain type. We may consider upgrade rules such that, if agent i comes to believe-in-the-limit that i and j can play a game, then i also comes to believe that i and j are of the same type. I will leave this proposal for future work.

6. Conclusions

Being or not being in a game, or being or not being able to fix a matrix, is in many cases a matter of luck. It is an empirical question, at least as much as stumbling on semantic ungroundedness or on a semantic paradox can also be an empirical matter (Kripke, 1975). Moreover, the circumstances that trigger ungroundedness can be knowledge, or simply belief, about other people's mental states. Sometimes beliefs about other people's mental states can lead to a scenario in which we do not know certain things about ourselves anymore — namely, we no longer know our payoffs, because the relevant matrix becomes underdetermined. And, in extreme cases, beliefs about other people's mental states can lead to a scenario in which the very possibility to interact game theoretically dissolves.

In this paper I suggested that we can mimic the search for a suitable matrix by means of a dynamic logic with enough resources to represent both the beliefs and the preferences of the potential players. Agents upgrade their beliefs as a result of (soft)-learning about the preferences of their partner. The sequence of models can, but need not, come to an end for pairs of Perfect Lovers/Haters, but not for mixed pairs. Failures to reach a fixed point in the sequence of upgrades can be due either to lack of coordination, or to deeper impossibility. If an infinite loop occurs, what oscillates forever is not a belief about how the actual world is, but the very *determination* of the actual world (a world where a game can take place). In any case, recall that, as we have seen in Figure 2, stability is not sufficient for players to be in a game. By Theorem 2, only homogeneous pairs can acquire common belief that they can play a game, regardless of whether they are engulfed by an infinite sequence of upgrades.

A word of caution. Suppose that the agents do find a world where a game can be played. This does not amount to saying that the players reached a Nash equilibrium; they may be still in doubt regarding which strategy profile is implemented, or even which game they are actually playing. In other words, by saying that agents believe that they can play a game, we guarantee that they agree on at least one entry of a possible matrix, and hence on a family of matrices (those with the same entry). If the two agents coincide in world w, this means that they have agreed on *some* matrix — namely, one that includes as one of its cells the outcome specified by world w. However, players need not be opinionated regarding which matrix they are dealing with, even less so regarding which entry on a particular matrix gets instantiated. In light of all this, the import of observing that the agents achieve common belief on the proposition that they can play a game is more subtle than what it may have seemed at first sight. In previous work I suggested that players try to coordinate their first order matrices by (implicitly) engaging in a second-order game. By coming to believe that they *can* play a (first-order) game, players acquire common belief on the set of available matrices, and hence they can proceed to play the relevant second-order coordination game with complete information.

The present results can be developed further in future research along two distinct lines. On the one hand, we may want to explore a general account of the task of finding the matrix of a game. From the players' point of view, the process of looking for game theoretic matrices may be recast as a series of model upgrades, even when there is no payoff ungroundedness.

On the other hand, dynamic operators for preferences and beliefs could be made explicit in the language. This would allow for the formulation of precise rules (so-called reduction axioms) that systematically translate formulas involving dynamic operators into equivalent static ones.[16] It should be noted that no axiomatization is currently known for the full logic developed in this paper. While the underlying doxastic component is standard (KD45), and static fragments of the preference language fall within well-studied axiomatizable systems, the framework introduced here essentially combines dynamic belief upgrade with limit notions such as belief-in-the-limit, which quantify over all possible future upgrade paths. Logics with this kind of semantic quantification over infinite dynamic processes are well known to pose substantial obstacles to standard axiomatization techniques. That said, several promising directions for axiomatization can be identified.

One natural strategy would be to isolate axiomatizable fragments of the language (such as the static belief-preference fragment, or the dynamic fragment without limit operators) and then study principled extensions.

A conceptually distinct avenue would be reduction-based. Rather than restricting the language, one could investigate whether suitable reduction axioms for dynamic operators can be formulated, translating dynamic formulas into an enriched but purely static language, in the tradition of dynamic epistemic logic. For the doxastic component, the belief operator B_i behaves as in standard KD45-based dynamic epistemic logic, and radical upgrades $\Uparrow\phi$ admit reduction principles analogous to those familiar from DEL; thus, formulas of the form $[\Uparrow\phi]B_i\psi$ can be translated into static formulas expressing belief restricted to the most plausible ϕ-worlds. The situation is more delicate for preference operators, however. In the present framework, preference operators do not have independent dynamics, but are endogenously induced from payoff values, which in turn depend on agents' beliefs about others' preferences. Since $\preceq_i$ is induced from payoff values, and payoff evaluations depend on beliefs about others' preferences, reduction axioms for $\langle \mathrm{Pref} \rangle_i \phi$ would

[16] See (van Benthem and Liu, 2007; van Ditmarsch, Hoek and Kooi, 2007) among others. For more recent work on reduction axioms in dynamic preference logic, see (Souza, Moreira and Vieira, 2021).

necessarily interact with the reduction of belief operators. In particular, a reduction axiom for $[\Uparrow\phi]B_i\psi$ would have to make explicit how the upgrade affects i's beliefs about j's preferences, and hence how the induced preference ordering over worlds is recalculated.

Finally, operators such as belief-in-the-limit cannot in general be eliminated by finite reduction axioms, since their semantics quantifies over all future upgrade sequences. Belief-in-the-limit could plausibly be treated via fixed-point constructions, for instance within a μ-calculus — style extension of the language, at the cost of moving to a stronger logical framework. Exploring these possibilities goes beyond the scope of the present paper, whose primary aim is representational rather than proof-theoretic. Nevertheless, they define a clear agenda for future work.

"I prefer what s/he prefers." But if s/he prefers the opposite of what *you* prefer, then you cannot prefer that: the sentence describes an impossible world. You may not *know* that at first, though you may eventually convince yourself that this is so, when you upgrade. Interestingly, your "coming to know/believe" that this is so is a piece of knowledge/belief on your model dynamics, rather than on a specific world — because there is no such world in the first place.

Acknowledgements. Previous versions of this paper were presented at the ASL-APA Meeting in January 2021 and at the Tsinghua Logic Salon at Tsinghua University in May 2025. I am grateful to Gil Sagi and Fenrong Liu for their kind invitations and for very helpful feedback.

References

Bicchieri, C, 2006, *The grammar of society: The nature and dynamics of social norms*, Cambridge University Press.

Bonanno, G., 2015, "Epistemic foundations of game theory", pages 411–450 in H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, and B. Kooi (eds.), *Handbook of logics for knowledge and belief*, College Publications.

Brandenburger, A., and H. J. Keisler, 2006, "An impossibility theorem on beliefs in games", *Studia Logica*, 84: 211–240.

Cresto, E., 2022, "Ungrounded payoffs: A tale of perfect love and hate", *Journal of Philosophy*, 6(119): 293–323. DOI: 10.5840/jphil2022119621

Dekel, E., and M. Siniscalchi, 2015, "Epistemic game theory", pages 619–702 in H. P. Young and S. Zamir (eds.), *Handbook of Game Theory with Economic Applications*, vol. 4, Elsevier.

Di Tillio, A., 2008, "Subjective expected utility in games", *SSRN Electronic Journal*, 3(3): 287–323. DOI: 10.2139/ssrn.1494893

Estlund, D., 1990, "Mutual benevolence and the theory of happiness", *The Journal of Philosophy*, 87(4): 187–204.

Fehr, E., and K. M. Schmidt, 1999, "A theory of fairness, competition, and cooperation", *The Quarterly Journal of Economics*, 114(3): 817–868.

Gul, F., and W. Pesendorfer, 2016, "Interdependent preference models as a theory of intentions", *Journal of Economic Theory*, 165: 179–218.

Kitcher, P., 1993, "The evolution of human altruism", *The Journal of Philosophy*, 90(10): 497–516.

Kitcher, P., 2010, "Varieties of altruism", *Economics and Philosophy*, 26(2): 121–148.

Kripke, S., 1975, "Outline of a theory of truth", *The Journal of Philosophy*, 72(19): 690–716.

Liang, Z., and J. Seligman, 2011, "The dynamics of peer pressure", pages 237–250 in H. van Ditmarsch, J. Lang, and S. Ju (eds.), *Logic, Rationality, and Interaction (LORI 2011)*, Lecture Notes in Computer Science, vol. 6953, Springer. DOI: 10.1007/978-3-642-24130-7_32

Liu, F., 2008, *Changing for the Better: Preference Dynamics and Agent Diversity*, ILLC Dissertation Series DS-2008-02.

Liu, F., 2009, "Diversity of agents and their interaction", *Journal of Logic, Language and Information*, 18(1): 23–53.

Liu, F., J. Seligman, and P. Girard, 2014, "Logical dynamics of belief change in the community", *Synthese*, 191: 2403–2431.

Pacuit, E., and O. Roy, 2016, "A dynamic analysis of interactive rationality", pages 187–206 in J. Redmond, O. Pombo Martins, and A. Nepomuceno Fernandez (eds.), *Epistemology, Knowledge and the Impact of Interaction*, Logic, Epistemology, and the Unity of Science, vol. 38, Springer.

Perea, A., 2012, *Epistemic Game Theory: Reasoning and Choice*, Cambridge University Press.

Rabin, M., 1993, "Incorporating fairness into game theory and economics", *American Economic Review*, 83: 1281–1302.

Seligman, J., F. Liu, and P. Girard, 2011, "Logic in the community", pages 178–190 in M. Banerjee and A. Seth (eds.), *Logic and its applications (ICLA 2011)*, Lecture Notes in Computer Science, vol. 6521, Springer. DOI: 10.1007/978-3-642-18026-2_15

Souza, M., A. Moreira, and R. Vieira, 2021, "Dynamic preference logic meets iterated belief change: Representation results and postulates characterization", *Theoretical Computer Science*, 872, 15–40.

Stalnaker, R., 1994, "On the evaluation of solution concepts", *Theory and Decision*, 37: 49–73.

van Benthem, J., and F. Liu, 2007, "Dynamic logic of preference upgrade", *Journal of Applied Non-Classical Logics*, 17(2): 157–182.

van Benthem, J., and F. Liu, 2016, "Deontic logic and changing preferences", Chapter 9 in D. Gabbay, J. Horty, X. Parent, R. van der Meyden, and L. van der Torre (eds.), *Handbook of Deontic Logic and Normative Systems*, vol. 2, College Publications.

van Benthem, J., and S. Smets, 2015, "Dynamic logics of belief change", pages 299–368 in H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, and B. Kooi (eds.), *Handbook of Logics for Knowledge and Belief*, College Publications.

van Benthem, J., S. van Otterloo, and O. Roy, 2006, "Preference logic, conditionals and solution concepts in games", pages 61–77 in H. Lagerlund, S. Lindström, and R. Sliwinski (eds.), *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, Uppsala Philosophical Studies.

van Ditmarsch, H., W. van der Hoek, and B. Kooi, 2007, *Dynamic Epistemic Logic*, Springer.

Eleonora Cresto
Universidad de San Andrés, Argentina
ecresto@udesa.edu.ar
https://orcid.org/0000-0003-1820-8179