**Ungrounded Payoffs**

**A Tale of Unconditional Love and Unrepentant Hate**


## Abstract

I explore a game theoretic analysis of social interactions in which each agent's well-being depends crucially on the well-being of another agent. As a result of this, payoffs are interdependent and cannot be fixed, and hence the overall assessment of particular courses of action becomes ungrounded. A paradigmatic example of this general phenomenon occurs when both players are 'reflective altruists', in a sense to be explained. I begin by making an analogy with semantic ungroundedness and semantic paradoxes, and then I show how to proceed in order to model such interactions successfully. I argue that we obtain a second order coordination game for subjective probabilities, in which agents try to settle on a single matrix. As we will see, the phenomenon highlights a number of interesting connections among the concepts of self-knowledge, common knowledge and common belief.


Word count: 5975 words without the Appendix

# Ungrounded Payoffs

## A Tale of Unconditional Love and Unrepentant Hate

### 1. Introduction

> *Calvin is five years old, and Ron, his father, has just started to teach him to play tennis. Ron is concerned about Calvin's self-esteem, so he always lets his son win. Calvin's enthusiasm makes Ron immensely happy.*
>
> *Time goes on; Calvin is now ten. They play every Sunday. Ron still lets Calvin win most of the time, though not always – he wants Calvin to learn how to handle eventual defeats. Regardless of the outcome, they both love the time they spend together.*
>
> *Eventually Calvin becomes a teenager; Ron at this point is slightly overweighed and not as athletic as he used to be. They still play, but not as often. Soon Ron discovers that no matter how hard he tries, he just cannot win anymore; he is profoundly proud of his son.*
>
> *A few more years passed by. They have agreed to play more regularly, as Ron needs to exercise (so says the doctor). Ron notices with certain satisfaction that after a bit of practice on his part they are even: sometimes Calvin wins, sometimes Ron wins.*
>
> *Then one day it hits him: now his son is letting him win.*[1]

This story can be said to speak about many different issues. But here I want to focus on one possible consequence, or side effect, of the situation our two characters stumbled upon at the very end. Let us conjecture a little bit how they might be feeling at that point. We may suspect that Calvin would rather prefer that his father does not realize that he is letting him win. On the other hand, on learning that Calvin *is* letting him win, Ron's possible small disappointment at his physical condition (he is not as fit as he thought he was) is likely to be more than compensated by his realization that his son wants to make him happy (it is crucial for the story that Ron understands that Calvin does not want his father to find out that his son is letting him win). Thus, it is likely that Ron would rather keep on acting as if he had not noticed any of this, in order not to embarrass his own son. Calvin may in turn realize what is going on, smile at it, and try to figure out what to do at that point – probably, keep on pretending...

Let me take this story as a motivational force to analyze game-theoretically a very peculiar phenomenon. The situation I have in mind is one in which two agents are such that each of them wants to do whatever it is that makes the other agent happy *in his or her own terms.* (This is an important qualification, to which I will go back later: it amounts to a

[1] I read this brief vignette some years ago in a newspaper; I would like to give proper credit to his author, but unfortunately I cannot find the reference. In any case, many details have been modified.

peculiar form of altruism, which I will contrast with what we may call '*paternalistic* altruism'). If we try to pursue a standard game theoretic analysis of this situation, we soon realize that we run into trouble.

## 2. Not your standard coordination game

Before going on, let me clarify what this phenomenon *is not*. At first blush, one might think that our story is similar to other coordination games. But it is not.[2]

To see the contrast clearly let me recall very briefly how a standard game-theoretical analysis of a coordination game go – just for the sake of perspicuity. Suppose Alina and Benicio are invited to a party, and they consider bringing either wine or pie, but not both. Neither Alina nor Benicio knows what the other is doing, and they cannot call each other to check, for whatever reason. Suppose they both have exactly the same preferences. The ideal situation for any of them is one in which they can make sure that once in the party they will have something to drink *and* eat; so if Alina brings wine, Benicio would rather bring pie, and vice-versa. We typically rely on utilities as a way of representing the agents' preferences, and as a result we end up having a matrix with a certain payoff structure. In this particular case we will have two pair of strategies in equilibrium. This matrix still does not tell Alina and Benicio what to do – so it does not solve everybody's life problems. But none of this is problematic for game theory**.**

| | Pie | Wine |
|---|---|---|
| Pie | 0, 0 | 10, 10 |
| Wine | 10, 10 | 0, 0 |

Now consider again Ron and Calvin – or actually, any two agents (call them Row and Column) who engage in interactions just like they do. To simplify things, suppose that, depending on the results of their joint actions, the two agents can obtain one of two possible payoffs: a maximum payoff of 10, or a minimum payoff of 0. Recall that our heroes are such that, no matter what actions they consider doing, each cell has to reflect the following:

First of all, whatever action Row performs, it has maximum payoff for Row if and only if, by doing this, his interaction with Column reports Column a payoff of 10. Thus, if Column's payoffs are as stated below (in black)…

[2] Actually, we will see later that there *is* a sense in which Calvin and Ron are playing a sort of coordination game, but not in the obvious way. More on this later.

| | Column does *A'* |
|---|---|
| Row does *A* | 10, 10 |

| | Column does *A'* |
|---|---|
| Row does *A* | 0, 0 |

… then Row will have a payoff of 10 in the first case, and 0 in the second case (in red). This is not problematic – yet.

The problem arises when *in addition*, whatever action Column performs has a maximum payoff if and only if by doing this, his interaction with Row reports Row a payoff of 10. Thus, once again, if Row's payoffs are as sated below (in black), Column will have 10 in the first case, 0 in the second case (in red).

| | Column does *A'* |
|---|---|
| Row does *A* | 10, 10 |

| | Column does *A'* |
|---|---|
| Row does *A* | 0, 0 |

In other words, we see that Row's utility function depends on Column's utility function. This may be innocuous as long as Column's function is fixed independently (as when Calvin was a little kid: he just wanted to win the match). But, for empirical reasons, it may well occur that Column's utility function is in turn dependent on Row's function. (Actually, as we will see in a moment, for a particular type of altruists what matters is what both Row and Column *believe* about each other's payoff)

It is an empirical question whether such interdependence arises or not. If it does arise, I will say that we stumbled upon a setting in which the payoffs are *ungrounded*. As opposed to Alina and Benicio's problem of finding out the best way to party, this is not a standard coordination game: rather, *there is no matrix to begin with*. I will explain with more detail why this is so in Section 4.

Technically, this is not a problem *for game theory*, given that game theoretical considerations begin only once we have a matrix. As opposed to that, here we are dealing with the problem of how to fix a matrix in the first place. We can think of it as a problem of modelling, which is logically previous to (and a pre-condition for) any possible game theoretical analysis. In any case, it is still a problem for game theoretic *representation* in a broad sense. At least at first blush, the impression is that we have stumbled upon a phenomenon that seems to be non-expressible in game theoretical terms, while on the other hand we do have the intuition that, in these and similar scenarios agents somehow can figure out how to act in a rational way (even in a strategically rational way). So the question is how

this riddle is to be solved. We shall see that the phenomenon has at least indirect relevance both for game theory and for (formal) epistemology, in a very general way.

Before going on, a few clarifications are in order.

First, bear in mind that "solving the modelling problem" does not mean "solving the game". I trust the two expression are distinct enough so that not confusion will arise, but it does not hurt to insist on this.

Second, it is clear that real life agents will always do *something*, even if what they do ca no longer be described as purposefully choosing a strategy. For example, they might just play tennis mechanistically without any clear purpose in mind. In this scenario, there is no longer a game in a theoretically interesting sense. But *at least sometimes* agents in situations as the one depicted above do ponder what to do and try to behave strategically. As long as this is the case, the conceptual problem remains.

Finally, I would like to emphasize that our initial story was only meant to be motivational; it was an excuse to develop a theoretical problem for a peculiar form of altruism. I will not try to work out all the details of Calvin and Ron's story, which involve complexities that go beyond the goals of the present paper. For one, a tennis match develops over time, so there is room to think that players are actually dealing with a continuum of decisions. By contrast, here I will focus on cases in which single simultaneous decisions are at stake, and so I will examine them under the guise of normal (or strategic) form, exclusively. In the light of this, if the tennis example looks inappropriate to you, consider any other case in which Calvin and Ron wonder how to make each other happy by way of doing something specific (say, they might ponder whether to bring a hand-made pie or a pie they buy at the local supermarket, or whatever.)

## 3. Types of ungroundedness. An analogy

The phenomenon of ungrounded payoffs is not exclusive of unconditional love, though this is the example I will use here to develop the problem formally (let us say it is just good karma). Consider two neighbors who hate each other profoundly. Their mutual hostilities have escalated across the years, although at this point none of them remembers exactly how it all started. They just want to make the other angry – they gain maximum payoff from their neighbor's misery. Suppose that, to comply with a court order, each one is being forced to build a portion of a common fence in their backyard. Suppose that for whatever reason they can only choose between building a tall fence, or a very short one. They do not care at all about the fence *per se*, but they do care about the effect the fence will have in the neighbor.

Each of them stops to ponder the possible effects of their actions. Regardless of what Column decides to do, if Row builds a low fence, Row might be perceived as willing to spy on Column's surroundings (which is good for Row, as this might consolidate a reputation for aggressive behavior); alternatively, he might be perceived as vulnerable at the possibility of being spied upon himself (a bad thing for Row, as this might be perceived as a sign of weakness). If Row builds a tall fence, on the other hand, Row might come across as having succeeded in protecting himself from Column's potential gossiping (which is good for Row), or, again, he might be perceived as weak (say, as not willing to spy himself on Column, or

as feeling anxious about Column's possible gossips – any of which is again bad for Row). The same reasoning applies to Column.

As it happens, Row needs Column's payoff in each case to fix his own, and vice-versa. Let us call them 'unconditional haters'. As we can see, not all cases of ungrounded payoff are born alike.

For a more dramatic example (or at least, for an example with a more interesting literary pedigree), recall Kakfa's version of Ulises and the Sirens. In Kafka's story Ulises had decided to cover his ears, just like the sailors, but the Sirens foresaw the trick and did not sing, so as to avoid a defeat. Ulises, however, anticipated this move, but acted as if he didn't know: he didn't want them (and the Gods) to learn that he was aware of the fact that they could beat him *with their silence*. At this point we could easily continue Kafka's story and conjecture that the Sirens had considered this possibility as well… Who has beaten whom, in the end?

For the sake of completeness, we can also think of the sorry situation in which an unconditional lover faces an unconditional hater. As it will become clear in a moment, this situation is even more recalcitrant, from a game theoretical point of view, than the case of two unconditional lovers facing each other, or the case of two unconditional haters.

To see the difference among various cases of payoff ungroundedness more neatly, let me draw an analogy between the interdependence of utility functions and the interdependency of semantic valuations. Recall that semantic values can be interdependent in such a way that ungroundedness arises; in some such cases, in addition (but not in all of them), we will also have a semantic paradox. Consider the following three pairs of sentences:

(a) Sentence (b) is true
(b) Sentence (a) is true

(a') Sentence (b') is false
(b') Sentence (a') is false

(a'') Sentence (b'') is true
(b'') Sentence (a'') is false

Suppose the logic is classic (only two truth-values, no gaps, no gluts, etc.). Then, for the first pair, semantic values are ungrounded but not paradoxical: sentences (a) and (b) can have stable truth-values in case they are both true. For sentences (a') and (b') to receive stable truth-values, on the other hand, they need to have opposite values: (a') is true if and only if (b') is false. Again, there is ungroundedness, but not paradoxicality. As for sentences (a'') and (b''), they simply cannot receive stable truth-values: if (a'') is true, then (b'') is true, which means that (a'') is false, which means that (b'') is false, which means in turn that (a'') is true, and this goes on forever.

Recall, moreover, that it can be an empirical matter whether we stumble upon ungroundedness or paradoxicality, as Kripke taught us long time ago. In Nixon's example (regarding his assertions about Watergate), what makes the situation paradoxical is a contingent fact; we had just bad luck (Kripke 1975, pp. 691-692).

As with semantic valuations, scenarios in which we find interdependent utility functions can sometimes reach stable payoffs. Indeed, in our examples from previous sections, with only

maximum and minimum values, we notice that there are two available fixed-point pairs of payoffs for unconditional lovers facing each other, and there are also two fixed-points pairs of payoffs for unconditional haters:

*Unconditional Love (coordination-type ungroundedness):*

| | Column does *A'* |
|---|---|
| Row does *A* | 10, 10 |

| | Column does *A'* |
|---|---|
| Row does *A* | 0, 0 |

*Unconditional Hate (conflict-type ungroundedness):*

| | Column does *A'* |
|---|---|
| Row does *A* | 0, 10 |

| | Column does *A'* |
|---|---|
| Row does *A* | 10, 0 |

As we can see, then, we have coordination-type ungroundedness, as well as conflict-type ungroundedness. Had we allowed for intermediate outcomes (that are neither maximally preferred nor maximally hated by the agents), we could obtain further fixed points, as is obvious. But this would complicate the analysis without any real conceptual gain, so in what follows I will keep on assuming that in these particular interactions agents only obtain best or worst outcomes.

On the other hand, there is no possible pair of payoffs that captures the result of an unconditional lover facing an unconditional hater:

*Unconditional Love vs. Unconditional Hate:*

| | Column does *A'* |
|---|---|
| Row does *A* | ?, ? |

As with its semantic counterpart, in this last case the interdependence of utility functions is such that payoffs are not only ungrounded, but impossible. We stumbled upon a paradoxical setting.

In this paper I seek to solve the modelling problem for non-paradoxical ungroundedness, and at least for now I will not attempt to deal with the paradoxical scenario. Moreover, I will focus on coordination-type ungroundedness (i.e., cases in which fixed points require equal payoffs). The analysis for conflict-type ungroundedness, however, is symmetric to the one I will offer today, so it is not hard to extend the proposal to cover this case as well.

As I have pointed out in the previous section, coordination-type ungroundedness is not akin to a standard coordination game, as there seems to be no matrix to begin with. Our next task is to fully understand why this is so, and which consequences follow from this.

## 4. Understanding the phenomenon. Types of altruism

Let us go back to our tennis match, and to the possible actions performed by altruistic agents (or 'lovers' – as opposed to 'haters'), to see if we can have a better grasp of what is going on. To begin with, notice that an action can be said to be altruistic in very different senses, and not all such senses lead to interdependent utility functions.

Consider a first level of altruism, which I will call '*Paternalistic Altruism'*. Each one wants *the other* to win – and this because each one just assumes that this is what the other wants. Virtually all research done by contemporary game theory on altruism concerns this interpretation of altruist behavior.

How shall we model this? Take the relevant strategies to be 'try to win'/ 'try to let the other win' (rather than, say, 'win' or 'lose' – as these are not under each agents' control). To simplify things, moreover, assume that Row and Column are sufficiently similar to each other in the relevant respects, so that if they choose the same strategy they will win and lose evenly in the long run, whereas if Row [Column] tries to let Column [Row] win while Column [Row] tries to win herself, then indeed Column [Row] wins. Then the payoffs are as follows (payoffs in the diagonal cells can be taken to be a middle point between the payoffs of winning and losing)**:**

| | Try to win | Try to let Row win |
|---|---|---|
| Try to win | 5, 5 | 0, 10 |
| Try to let Column win | 10, 0 | 5, 5 |

'Try to let the other win' is obviously the dominant action. So far, this is a completely non-problematic scenario.

Now consider a second level of altruism – what we might dub '*Sensitive Altruism*'. Now each agent wants the other to do what s/he really wants to do, whatever that might be. From Row's point of view, if Row thinks that Column wants to win, Row will be happy to let him win; by contrast, if Row detects that Column wants to let Row win, Row will be

happy to win. The same goes for Column. What they still do not know – and this is crucial – is that the situation is symmetric for the two. Then again we have a non-problematic matrix (and one that resembles the wine-and-pie scenario):

| | Try to win | Try to let Row win |
|---|---|---|
| Try to win | 0, 0 | 10, 10 |
| Try to let Column win | 10, 10 | 0, 0 |

Here we suppose each one assumes that whatever the other does is what s/he really wants to do. We then have a standard coordination game with two *equilibria* points. Notice that we can only model this situation from the outside, as it were – neither Row nor Column are the modelers.

Finally, consider a third type of altruism, or '*Reflective Altruism'*. Each agent has just grasped what is going on – say, they have become the modelers themselves. A somewhat paradoxical (though ultimately inexact) way of putting it would be to say that having a perfect access to the putative structure of the pretended matrix results in the fact that the players no longer have a matrix. More precisely, we will say that agent 1 is a (full) reflective altruist (with respect to agent 2) if, for all profiles, agent 1 has maximum payoff iff agent 1 believes that agent 2 has maximum payoff; iff agent 1 believes that agent 2 believes that agent 1 has maximum payoff; iff 1 believes that 2 believes that 1 believes that 2 has maximum payoff; iff… etc. So defined, this is what we might call "subjective" reflective altruism, given that one agent's utility is determined by *her beliefs* in her partner's gain. In our simplified analysis, moreover, recall that not having maximum payoff amounts to having minimum payoff.[3] (For a more precise, formal definition, see the **Appendix**).

How does the problem arise, exactly? Row reasons thus: "if Column thinks I want him to win, she will try to win; otherwise, if she thinks I want to win, she will let me win. In other words, Column will try to win (i.e., Column thinks she will maximize her payoff by trying to win) iff she thinks that is what I want from her." An analogous reasoning goes for Column.

The problem is that many different situations are compatible with this biconditional. For one, Row might conjecture that Column will rather not try to win if Row tries (given that if Row tries to win, a best response for an altruist would be to let him do whatever he is trying to do). In this case we would be back at the previous matrix – that of the sensitive altruist.

But of course this need not be the case. Row might rather think that Column wants both agents to try to win at the same time; for example, he might reason that that is the best way to avoid paternalistic traps.

Alternatively, he might think that Column will be happy whenever both agents chooses the same strategy, whatever that will be (either try to win or try to let the other win).

[3] If this assumption is lifted the analysis gets more convoluted, but the central modelling problem remains unchanged, and the main solution I propose in this paper will still be effective. So for the moment I will keep the analysis as simple as possible.

There are other possibilities as well, of course. Actually, as we have two fixed points (either (10, 10) or (0, 0)), *any* matrix with payoffs (10, 10) or (0, 0) can do the trick. Given that our putative matrix will have four cells (i.e., two possible actions for each player), this gives us 16 possible matrices, from all (10, 10) points to all (0, 0) points.

Let me say once again that, as we have already noticed in the last section, even though payoffs are ungrounded, players are not locked into a paradoxical loop. It is a case of ungroundedness without paradoxicality, because pairs of payoffs (10, 10) and (0, 0) are actually unproblematic. In other words, any one of the 16 possible matrices would be perfectly OK. Only, which one is the one? Unfortunately, there is no clear pair of actions that yields a unique set of fixed points for any possible level of reflection (or iterated belief). *Our modeling problem can then be described by saying that the situation is such that we are forced to oscillate between the sixteen different matrices.*[4]

How can we go from here? A possible way out is to assume that agents conjecture that their partners have particular payoffs for particular profiles, and then they let their own payoffs to accommodate. *If they are lucky*, some matrix will settle in – as long as all results have payoffs of the form (*x, x*), for some real number x that represents the same level of preference.[5] In the next section I will discuss a possible way to reach such fixed-point outcomes through probabilities.

Let me close this section with some observations on the relation between reflective altruism and belief. Reflective altruism is indeed a strong concept, and we may wonder, for example, whether a pair of reflective altruists will always believe that they are both reflective altruists, or whether their being reflective altruists entails that there is common belief between them.

Perhaps somewhat surprisingly (given that we have infinite iterations), a pair of (subjective) reflective altruists need not have shared belief of their joint condition, unless we add the assumption that each agent is also *aware* of her being a reflective altruist. Even with this assumption in place, there need not be common belief between the two agents regarding their condition. For a formal proof of these claims see the **Appendix**, where I present a suitable formalism to account for these observations (**Propositions 1, 2 and 3**).

This is interesting, I take it, insofar as it is rather natural to think that the problem arises because the two agents have 'too much knowledge' – or at least 'too much belief: i.e., because they know, or believe, that they are both (fully) reflective altruists. But this is wrong: a pair of agents need not have common believe in the fact that they are both reflective altruists to be in trouble. It is enough for them *to be* reflective altruists, even if they do not believe so.

---

[4] Someone might contend that the actual courses of action open to the agent (trying to win the game, baking a pie, or whatever) are no longer important at this point, because, ultimately, agents are playing a different game – a *pretense* game. In a pretense game each agent signals his happiness to his partner (or maybe *fakes* it), to make him/her happy. We might think that this interpretation allows us a simpler way of avoiding the conceptual trap of not having a clearly defined matrix to begin with, and hence that the modelling problem is easily solved. On a closer look, however, this is not so. First, it is not obvious whether they will believe each other. If Row does not believe Column, no matter how much Row fakes happiness, she will be actually miserable; the same goes for Column. In addition, there is still the question of what to do *within the original setting*. If anything, they will seek to convince their partner of their happiness through a specific action. So strategies now are, for instance, "signaling happiness while trying to win/trying to let the other win". And this brings us back to our original set of profiles.

[5] Compare this situation with that of unconditional haters, in which only pairs of the form (*x, y*) will do, where *x* and *y* are values representing the opposite extremes of their preferences.

Nonetheless, recall that undgroundedness is triggered by knowledge, or even simple belief, *about other people's mental states*. Thus a Sensitive Altruist will immediately become a *Reflective* Altruist if she knows, or believes, that her partner is a Reflective Altruist. An interesting moral of this situation is that sometimes knowing (or believing) too much about other people's mental states can lead to a scenario in which we no longer know certain things *about ourselves* anymore – namely, we no longer know our own payoffs.

## 5. Altruism, probabilities and second order games

Our problem will unravel once we realize that agents might have probabilities *over pairs of payoffs*. This is not tantamount to their having probabilities for each other's actions. Rather, this will help us identify their personal probabilities over the sixteen possible matrices, and to determine, eventually, whether they believe some matrix or other objectively corresponds to the game they are actually playing.

Let '$P_R$' be Row's subjective probability function, and '$P_C$' Column's function. Then, for any profile *a*, agents can conjecture how probable it is that their partner has it:

- $P_R$ (the payoff for Column in *a* is 10) = *x*

and

- $P_C$ (the payoff for Row in *a* is 10) = *x*

The probability Row assigns to each particular possible matrix (out of the 16) is then a combination of the aforementioned probabilities for each of the four profiles. (See the **Appendix**).

Notice that actors at this stage are playing *the second order game of trying to find a (first order) matrix*. If they are lucky, some matrix will finally settle. It is a particular coordination game: a game in which agents try to coordinate their subjective probabilities. We can think of it as having a 16 times 16 (second order) matrix (for any first order matrix with four profiles), in which both players obtain maximum payoff in the diagonal cells – that is to say, if and only if they reach an outcome in which they both believe they have the same matrix.

It is easy to see that Row maximizes her expected utility by choosing the matrix she deems it is more likely to be picked by Column (given that highest payoffs are in the diagonal). An analogous reasoning goes for Column.[6] In case more than one strategy maximizes her expected utility (if one or more profiles in the first order game have equal probability of yielding any of the two possible payoffs) we assume some secondary mechanism to break ties. We will obtain that the first order matrix 'selected' by Row (so to speak) will coincide with the one 'selected' by Column if we have, for all profiles *a* in the first order game:

$P_R$ (the payoff for Column in *a* is 10) $\geq 0.5$ iff
$P_C$ (the payoff for Row in *a* is 10) $\geq 0.5$

[6] Recall that, within the first order matrix, Row does not have probabilities *for Column's strategies*. The probabilities we have been considering so far refer to the payoffs associated with the four different possible profiles.

Notice that, if they reach a Nash equilibrium in the second order matrix, the first order matrix they select need not be a 'nice' one, in any sense. Solving the modelling problem is not to be confused with achieving a high payoff in the first order game. To illustrate the difference with an extreme case, suppose that, for both agents, and for all four profiles, the probability that the payoff for Column [Row] in a profile is 10 equals 0. Then the two agents will have probability one for a matrix in which the four profiles have pairs of payoff (0, 0). They will maximize their expected utility of the second order game by choosing this matrix – which means essentially that whatever they do within the first order scenario, that will bring them unhappiness. Still, our modelling problem is solved.

What if players do not reach a Nash equilibrium in their second order game? Then we are left again without a unique first order matrix, and the possibility of a first order game-theoretic representation fails. In the light of all this, it is not hard to see that a pair of (full) reflective altruists can have a matrix (and hence solve the first order modelling problem) iff they have common knowledge of its payoff structure (see the **Appendix**, Proposition 5).

Let me end this section with some observations on the relation between probabilities and beliefs. So far we have seen that attributing probabilities to payoffs is enough to determine the relevant elements of a second order game. But there is a further consequence of this procedure, which we haven't mentioned yet. Once a probability assignment allows Column to choose a matrix, we actually obtain (by way of the very biconditional embedded in the definition of a Reflective Altruist) that Column *believes* that Row has certain payoffs for particular profiles.[7] Then the whole hierarchy of beliefs can be built bottom up. This raises the interesting question of what kind of justification (if any) agents achieve by proceeding thus. This problem largely exceeds the scope of the paper, so let me just hint at a few ideas.

Someone might contend that the direction of the justification goes top-down, whereas the generation of the hierarchy works bottom-up. Let me put it differently. Let '*p*' and '*q*' refer to 'Row/Column's payoff in profile *a* is 10', respectively. We could argue that Row's belief that *p* is justified (to the extent that it is) *because* Row believes that Column believes that *q*, and this in turn *because* Row believes Column believes that Row believes that Column believes that *p*, etc. – we have an infinite chain of justifications.[8] However, the present setting seems to tell us that Row's believing that *p causes* his believing that Column believes that *q*, which *causes* his believing that Column believes that Row believes that Column believes that *p*, etc. So the first belief of the chain is not justified before actually generating the whole sequence. We seem to have stumbled upon some kind of bootstrapping justification.

More modesty, we could reject the claim that justification in this case only flows top-down. At least for subjective Bayesians, agents are entitled to having the probabilities they have (as long as they are not formally incoherent), and rational agents are required to

[7] This is not tantamount to the so-called Lockean thesis on belief. It is not that an agent beliefs in certain statement because it has probability higher than a certain threshold (where here the threshold would be set at 0.5), but because the statement describes features of a strategy (a matrix) that maximizes her chances of having maximum payoff. (Actually, if Row's probability for 'the payoff for Column in (a) is 10' is 0.5, whether Row believes that the payoff for Column is 10 or not will depend on the tie break mechanism).

[8] Notice that 'justification' here should be understood in a subjective sense: after all, Column can be wrong. This need not be a problem for the present argument, as long as we are clear on what we mean.

maximize their expected utility (as our agents are doing while playing their second order game). This seems an impeccable way of getting oneself in a position of forming a belief.

A purist may still insist that by doing so our agent might well be able to generate the belief that triggers the whole sequence, but that that first belief is still not justified: utility maximization is not the right kind of reason to acquire a belief. If we also endorse the claim that bootstrapping justification can never be right, then we are left with the view that the first belief of the hierarchy was not rationally acquired – period. Were we to adopt this standpoint, we may conclude that our modelling problem can only be solved for not perfectly rational agents, i.e., for agents that sometimes depart from rational belief formation. I am prepared to bite the bullet.

## Conclusions

I have argued that 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. The circumstances that trigger undgroundedness, in the case of utility functions, can be knowledge, or simply belief, about other people's mental states. An interesting moral of this situation is that sometimes knowing too much 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.

I have also argued that non-paradoxical ungroundedness regarding utility functions can be dealt with if we indulge in the (modest) assumption that agents attribute probabilities to payoffs, and thus to matrices. The result is that we move to a second order game, in which agents try to coordinate their subjective probabilities. If the players reach a Nash Equilibrium, then they succeed in fixing the matrix for their first order game. If they do not, then the first order game-theoretic representation fails.

## References


Aumann, R and A. Brandenburger (1995). "Epistemic Conditions for Nash Equilibrium". *Econometrica*, 63 (5): 1161-1180.

Bonanno, G. (2015). "Epistemic foundations of game theory." In H. van Ditmarsch, J.Y. Halpern, W. van der Hoek and B. Kooi (eds), *Handbook of Logics for Knowledge and Belief*, College Publications, pp. 411–450.

Fagin, R., J.Y. Halpern, Y. Moses, and M. Vardi (1995). *Reasoning About Knowledge*. Cambridge, Mass.: MIT Press.

Kripke, S. (1975). "Outline of a Theory of Truth". *The Journal of Philosophy*, 72 (19): 690-716.

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

## Appendix

Let *a* be a profile, and let sub-indices 1 and 2 refer to two different agents. Let $\pi_i(a)$ be agent *i*'s payoff in profile *a*. Consider:

$\pi_1(a)=10$ iff $B_1\pi_2(a) = 10$
iff $B_1B_2\pi_1(a) = 10$

$\pi_2(a) = 10$ iff $B_2\pi_1(a) = 10$
iff $B_2B_1\pi_2(a) = 10$

This is level-1 Reflective Altruism for profile *a* and agents 1 and 2, respectively, which we abbreviate as Level-1-$[RA_{i,a}]$. For what we may call Objective Reflective Altruism, just replace the *B*-operator by a *K*-operator.[9] If no further clarification is given, I will assume we are always dealing with Subjective RA.[10]

Recall that we are assuming that agents only have payoffs 0 or 10; hence for any profile *a*, 'not-$\pi_i(a)=10$' just amounts to '$\pi_i(a)=0$'. Notice, moreover, that a reflective altruist is always defined *in relation to someone else*: he is someone who thinks his partner is *also* a reflective altruist (otherwise, he would fix his own payoff by adjusting it to whatever he thinks his partner happens to have). The consequence is that reflective altruists are opinionated on their partner's payoffs, meaning that if Row does not believe Column has 10 (for a given profile *a*), then she believes Column has 0. In the light of this, Level-1 $[RA_{i,a}]$, for some partner *j*, can be re-written thus:

$$(\pi_i(a)=10 \;\&\; B_i\pi_j(a)=10 \;\&\; B_iB_j\pi_i(a)=10) \vee (\pi_i(a)=0 \;\&\; B_i\pi_j(a)=0 \;\&\; B_iB_j\pi_i(a)=0)$$

Level-*k* $[RA_{i,a}]$ just adds *k* conjuncts to each of the two disjuncts, where each new conjunct incorporates a new iteration of belief operators $B_i$ and $B_j$, in the obvious manner.

For a given profile *a*, let an 'optimistic' (respectively, 'pessimistic') reflective altruist up to level-*k* be someone for whom the first (respectively, the second) disjunct is verified. Moreover, let a *full* optimistic [resp., pessimistic] reflective altruist for profile *a* be one who has the whole, infinite hierarchy of beliefs. More precisely, let $[RA_{i,a}^{+}]$ stand for "*i* is a (full) optimistic reflective altruist, for profile *a*" (with respect to some partner *j*), and $[RA_{i,a}^{--}]$ for "*i* is a (full) pessimistic reflective altruist, for profile *a*" (and some partner *j*). Then, for agents 1 and 2:

$[RA_{1,a}^{+}] = \pi_1(a)=10 \;\&\; B_1[RA_{2,a}^{+}]$
$[RA_{2,a}^{+}] = \pi_2(a)=10 \;\&\; B_2[RA_{1,a}^{+}]$

and

[9] In the present context 'knowledge' should be understood merely as 'true belief'.

[10] Subjective Reflective Altruism allows for cases in which one of the agents is a bona fide Reflective Altruist while the other has a fixed, independent payoff for every profile. The modelling problem for such scenarios does not add any essential novelties, but it would be cumbersome (and space consuming) to work out the details here, so I will avoid them for now.

$[RA_{1,a}^{--}] = \pi_1(a)=0 \;\&\; B_1[RA_{2,a}^{--}]$
$[RA_{2,a}^{--}] = \pi_2(a)=0 \;\&\; B_2[RA_{1,a}^{--}]$

Now we can define more precisely what it is for an agent to be a (full) Reflective Altruist $[RA_i]$ (with respect to some partner *j*):

**Definition**
$[RA_i]$ = for every profile, either $[RA_i^{+}]$ or $[RA_i^{--}]$

Let 'shared (or mutual) belief' and 'common belief' have their standard definitions.[11] Thus agents 1 and 2 have shared belief in the fact that they are both full reflective altruists if it is true both that $B_1([RA_1] \;\&\; [RA_2])$, and that $B_2([RA_1] \;\&\; [RA_2])$. Whereas in order for them to have common belief in '$[RA_1] \;\&\; [RA_2]$' it should be true that $B_1([RA_1] \;\&\; [RA_2]) \;\&\; B_2B_1([RA_1] \;\&\; [RA_2]) \;\&\; B_1B_2B_1([RA_1] \;\&\; [RA_2]) \;\&\; \ldots$

**Proposition 1**:
If a pair of agents truly believes that each of them is a (full) reflective altruist, they will have shared belief of the fact that they are both reflective altruists, i.e.:
$([RA_1] \;\&\; [RA_2] \;\&\; B_1[RA_1] \;\&\; B_2[RA_2]) \Rightarrow B_1([RA_1] \;\&\; [RA_2]) \;\&\; B_2([RA_1] \;\&\; [RA_2])$

**Proposition 2**:
In order for a pair of agents to have shared belief in the claim that they are both reflective altruists, each agent needs to be a reflective altruist *and believe so*. In other words, the contrapositive of Proposition 1 also holds (so we actually have a biconditional):
$B_1([RA_1] \;\&\; [RA_2]) \;\&\; B_2([RA_1] \;\&\; [RA_2]) \Rightarrow ([RA_1] \;\&\; [RA_2] \;\&\; B_1[RA_1] \;\&\; B_2[RA_2])$

**Proposition 3**:
Shared belief in reflective altruism does not entail common belief.[12]

**Proof of Proposition 1**:
It is easy to see that $[RA_1] \Rightarrow B_1[RA_2]$:

$[RA_1] \Rightarrow$
$\Rightarrow$ For all profiles *p*: $(\pi_1(p)=10 \;\&\; B_1[RA_{2,p}^{+}]) \vee (\pi_1(p)=0 \;\&\; B_1[RA_{2,p}^{--}])$ (by definition)

[11] *Cf.* for example Fagin et al. (1995), or Bonanno (2015).

[12] As we will see, the reason is that we cannot guarantee that agent 1 believes that agent 2 believes to be *himself* a reflective altruist.

$\Rightarrow$ For any instance *a* of *p*: $B_1(\pi_2(a)=10 \ \& \ B_2[RA_{1,a}^{+}]) \vee B_1(\pi_2(a)=0 \ \& \ B_2[RA_{1,p}^{--}])$ (as this follows from both disjuncts)
$\Rightarrow$ For all *p*: $B_1((\pi_2(p)=10 \ \& \ B_2[RA_{1,p}^{+}]) \vee (\pi_2(p)=0 \ \& \ B_2[RA_{1,p}^{--}]))$ (generalization, plus standard behavior of the B-operator)
$\Rightarrow B_1$ (for all *p*: $(\pi_2(p)=10 \ \& \ B_2[RA_{1,p}^{+}]) \vee (\pi_2(p)=0 \ \& \ B_2[RA_{1,p}^{--}])$)
$\Rightarrow B_1[RA_2]$.[13]

Analogously, $[RA_2] \Rightarrow B_2[RA_1]$. Moreover, if we assume that agents are aware of the fact that they are themselves reflective altruists we will also have $[RA_1] \Rightarrow B_1[RA_1]$ and $[RA_2] \Rightarrow B_2[RA_2]$.[14]

Putting all this together, we obtain shared belief on $[RA_1]$ & $[RA_2]$. ■

**Proof of Propositions 2 and 3**:
We will show suitable counterexamples by building a couple of semantic models. We will rely on the standard Kripke setting for the B operator – a serial, transitive and Euclidean frame, corresponding to the KD45 system.

As usual, a model is a tuple $<W, R_1, R_2, L>$, where W is a set of worlds, $R_1$, $R_2$ are accessibility relations for agents 1 and 2, respectively, and *L* is a suitable modal language. Worlds will be represented by an indexed pair (with possible indices $i = 1, 2,\ldots$) consisting of a profile and a pair of payoffs, for 0 and 10 as the only possible payoffs; two worlds can capture the same pair and still be distinct.[15] So, for example, we may have worlds such as $(a, (10, 10))_1$; $(a, (10, 10))_2$; $(a, (10, 0))_3$; $(b, (0, 0))_4$, etc. Of course, there should be at least as many worlds as profiles.

We will indulge in worlds whose pairs of payoffs do not constitute a fixed point for pairs of reflective altruists. This is what we need in order to represent the possible failure of the (first order) game-theoretic representation. Notice that worlds with pairs of payoffs (10, 0) or (0, 10) are not logically impossible (in the sense a '*p* & not-*p*-world' is impossible); they are just impossible scenarios *for pairs of reflective altruists who are aware of their being so*. For our present purposes, partial descriptions of models will do.

Figures 1 and 2 below represent partial diagrams of two different possible models, *M* and *M**, with set of worlds *W* and *W**. Each of them shows the relations among a subset of worlds of *W* and *W**, respectively, for (part of) a single profile *a* (there is no need to assume all worlds for profile *a* have been included in each diagram). As there is no room for

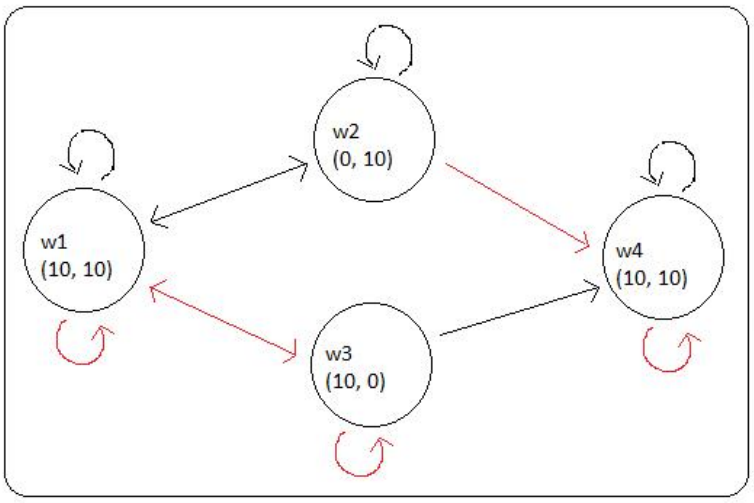
[13] A note on the penultimate step. Any sensible semantic model will have at least as many worlds as profiles – and, in any case, each world should be correlated with a single profile. A sentence with a B-operator is true in a given world *w* if the sentence within its scope is true in all worlds that relate to *w*. Moving the quantifier on profiles inside the scope of the B-operator does not have any effect on the set of worlds that relate to a given world.

[14] This is not to be confused with the stronger claim that if an agent has a certain payoff, then she believes she has it. A self-conscious reflective altruist may be unsure as to whether she is an optimist or a pessimist reflective altruist.

[15] It is standard practice in the literature that worlds need not *identify* with profiles. See for example Stalnaker (1994).

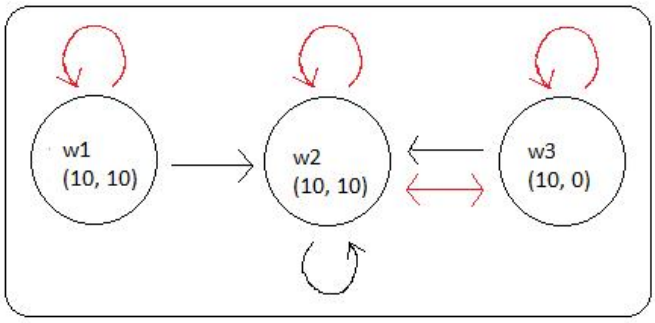
confusion, I have dropped the reference to profile *a*, and have just written the corresponding pair of payoffs inside each world. Black and red arrows refer to accessibility relations $R_1$ and $R_2$, respectively.

Figure 1 – A partial specification of *M*, for (part of) profile *a*.




Consider w1 from Figure 1. Both $[RA_1]$ and $[RA_2]$ are true in w1, but $B_1[RA_1]$ and $B_2[RA_2]$ are not, and as a result there is no shared belief of the fact that the two agents are reflective altruists.[16] ■

Figure 2– A partial specification of *M**, for (part of) profile *a*.




On the other hand, in w1 from Figure 2 the agents have shared knowledge of ($[RA_1]$ & $[RA_2]$), but $B_1B_2[RA_2]$ is false, so it is easy to see that there is no common belief. ■

[16] Incidentally, notice that in w2 agent 2 falsely believes that 1 is an (optimistic) reflective altruist. *Mutatis mutandis* for w3 and agent 1. Again, no shared belief is possible in w2 or w3.

**Reflective Altruists and Probabilities**

We enrich our model with finitely additive probability measures $P_1$ and $P_2$ on sets of worlds. Note that, given our earlier assumptions, for each profile $a$, $\pi_i(a)$ partitions $W$ into two sets. Thus we can define a random variable $V_a$ for each profile $a$, with values 0 and 10. Moreover, we assume that for any two profiles $a$ and $b$, $V_a$ and $V_b$ are probabilistically independent.

Consider now a first order game $g$ with profiles $a_1,\ldots a_k$. The probability that the game has a particular matrix $m$, for agent $i$, is:

$P_i(m) = \prod_{j=1}^{k} P_i(V_{a_j} = x)$

(Thus, for a 4-profile game, the sum of all 16 possible matrices adds to 1, of course).

Next, define a game $g'$ with profiles $(m,m')$, where $m$ and $m'$ are intuitively the strategies (i.e., first-order matrices) available to agents 1 and 2, respectively. Payoff are such that $\pi_1(m,m') = \pi_2(m,m')$ is maximum iff $m = m'$. From an initial first order game with $n$ profiles and two possible payoffs we obtain $m_k$ strategies for each agent, where $k = 2^n$, as is obvious. Agent $i$ then maximizes his expected utility by choosing a strategy $m$ such that $P_i(m) \geq P_i(m_k)$, for any other $m_k$. As is customary, I will assume agents know which strategy they pick.

**Proposition 4:**
For every profile $a$: $B_i(\pi_j(a)=10)$ iff profile $a$ has payoff 10 in the matrix selected by $i$.

Straightforward from the definition of Reflective Altruism. ■

**Proposition 5**:
A pair of (full) reflective altruists can have a matrix (and hence the first order modelling problem is solved) iff they have common knowledge of the payoff structure of the matrix.[17]

**Proof of Proposition 5**:
The proof is straightforward, if we assume the standard semantics for the K operator. Let $\pi_{i,j}(a) = (\cdot\,,\cdot)$ be the pair of payoff for agents $i$ and $j$ in profile $a$. If a pair of full reflective altruists has a (common, objective) matrix, then for every profile $a$: $\pi_{i,j}(a)=(10,10)$ iff $B_i(\pi_j(a)=10)$, for any (possibly identical) $i$ and $j$ (by proposition 4). Hence for two agents $i$ and $j$, and any profile $a$, $\pi_{i,j}(a)=(10,10) \Rightarrow K_i(\pi_{i,j}(a)=(10,10)) \;\&\; K_j(\pi_{i,j}(a)=(10,10))$, so the two agents have shared knowledge of their matrix. Consider now agents 1 and 2. By definition of a reflective altruist, $K_1(\pi_2(a)=10) \Rightarrow K_1B_2(\pi_1(a)=10)$, whereas $K_1(\pi_2(a)=10)$ & $\pi_{1,2}(a)=(10,10) \Rightarrow K_1K_2(\pi_1(a)=10)$. In addition, we have both $K_1B_2(\pi_2(a)=10)$ and $K_1(\pi_2(a)=10)$, and hence $K_1K_2(\pi_2(a)=10)$. Thus $K_1K_2(\pi_{1,2}(a)=(10,10))$. Moreover, by axiom 4 we also have $K_1K_1(\pi_{1,2}(a)=(10,10))$. Hence $K_1\,(K_1(\pi_{1,2}(a)=(10,10)) \;\&\; K_2(\pi_{1,2}(a)=(10,10)))$. An analogous result follows for agent 2, and it is easy to see that the procedure can be iterated indefinitely. (The full proof goes by induction on the number of iterations on $K_i$). ■

[17] This does not imply that agents have common knowledge *on their second order matrix*. Cf. Aumann and Branderburger (1995).