Frames and games: intensionality and equilibrium selection

István Aranyosi (Bilkent University)

-- Final version forthcoming in *Erkenntnis* --

There is by now a venerable tradition in the literature on rational decision and economic behavior that focuses on so-called *framing effects*, first discussed in Amos Tversky and Daniel Kahneman's influential 1981 paper (Tversky and Kahneman 1981).[1] The core idea is that the way a decision problem is framed influences how agents deal with it in terms of decisions they make. Agents will choose differently depending on how their options are presented, even though, mathematically, the situations are equivalent. It has been customary for theorists to characterize it as a cognitive bias since it violates basic requirements of rationality. Some theorists have characterized it is an instance of the well-known phenomenon of extensionality violations that are frequently discussed in logic and in the philosophy of language (Arrow 1982, Bourgeois-Gironde and Giraud 2009, Ahn and Ergin 2010). The goal of this paper is to further connect the decision-theoretic ideas and those from the philosophy of language along the lines of this latter characterization of framing effects as extensionality violations; more precisely, a notion of what I will dub *Intensional Nash Equilibrium* will be distilled and presented as a useful equilibrium concept in game theoretic contexts. I will rely not on the Tversky-Kahneman extensionalist[2] lineage of the analysis of the framing effect, but rather on the intensional decision theory approach championed over the years by Frederic Schick (1991, 1992, 1997, 2003),[3] partly because his focus is explicitly on linguistic ambiguity, which, though I will also criticize it, is the right approximate domain to consider in the context of extensionality violations, and partly because I do sympathize with his alternative to the Tversky-Kahneman approach, namely, the idea that we need to

[1] Among others, Frisch1993, Kühberger 1998, Levin *et al* 1998, Gold and List 2004.
[2] Namely, where extensionality is assumed as an axiom of rational decision, hence violations of it are always instances of irrational behavior.
[3] And continued in Bermudez 2008, 2009, 2020. Alternative versions of intensional game theory are in Bacharach 1999, 2001, 2006 and Sugden 1993, 2000, 2003.

incorporate *understandings* into decision theory instead of simply judging victims of framing effects as irrational. Schick's point is that we need to consider the way agents see or understand the choices they have, and only then judge their choices. Understandings are, therefore, ways in which a decision appears to the agents involved in it. Violations of extensionality (or "invariance" in Tversky and Kahneman's terminology) are, therefore, not always indications of irrationality.

In section 1, I introduce and discuss a game theoretic version of extensionality violation in experimental setting discovered in the late 1960s –the case of the decomposed Prisoner's Dilemma game (PD). My main goal will be to inquire into how we could make sense of games as having Fregean senses and reference, which, in turn, are the conceptual skeleton on which extensionality violations are typically analyzed. In section 2, I continue focusing on PD as analyzed by Schick (2003) via his idea of "solving it" using intensionality as linguistic ambiguity and I offer some critical remarks on it. In section 3, I integrate my proposal of thinking about games along Fregean lines with similar suggestions in the game theory and experimental psychology literature. Section 4 is dedicated to the idea of an intensional Nash equilibrium concept; I exemplify its possible applications in other games than the PD. I conclude with a section on player identity as a source of intensionality and with a few thoughts about how the game theoretic application of framing as intensionality (rather than as mere cognitive bias) is a move that is less radical than some other well-known alternatives to classical decision theory and game theory, such as behavioral economics[4] or evolutionary game theory.

1. Sense, reference, and opacity in games

It was Gottlob Frege ([1892] 1980) who, almost hundred thirty years ago, first pointed out that there is more to meaning than reference (i.e., what a term designates) and offered a detailed analysis of the idea of this extra ingredient to semantics, which he called *Sinn* in

[4] I use "behavioral economics" here as synonymous with purely posteriori experimental psychological study of economic behavior and the consequent generalization and modelling, as opposed to an a priori modelling of normative game- and decision theory. By "less radical", then, I mean closer to this classical, normative, a priori modelling.

German, translated as "sense". Frege's starting point was a puzzle about identity, namely, the question: how is it possible to have informative, that is, cognitively significant, identity statements? How are 'The Morning Star = The Morning Star' and 'The Morning Star = The Evening Star' both expressing identity, when the former is trivial, whilst the latter expresses the outcome of an important empirical discovery?[5] Frege's line of thought was that there can only be three options when it comes to the question of what the relata of an identity relation are, namely

(i) that they are the objects referred to by the names that flank the identity symbol (their reference),

(ii) that they are the names themselves, and

(iii) that they are something distinct from both the reference and the names,

but options (i) and (ii) can safely be discarded since the former would render all identity statements trivial (at the level of reference all identity is self-identity of the sort 'a = a') and the latter would also lead to trivialization, because names qua names are simply arbitrary symbols, hence 'a = a' would not differ from 'a = b'.

The further constituent of meaning is what Frege calls "the sense of the sign, wherein the mode of presentation is contained" ([1892] 1948: 210). Informative identities, such as 'The Evening Star = The Morning Star', are, therefore, cases in which we have the same reference (planet Venus) but distinct senses (*the celestial object visible in the evening* and *the celestial object visible in the morning*). A sense of a proper name, according to Frege, is a mode of presentation of its reference. One and the same reference could be designated by different proper names with different corresponding senses. Frege goes on to analyze a host of other phrase types in terms of his sense/reference distinction. Thus, the sense of a sentence is a thought, whereas its reference is a truth value (*The True* or *The False*). The reference of a subordinate clause in a propositional attitude report, such as 'J believes *that P*', 'J desires *that P',* or 'J thinks *that P'* is, according to Frege, its customary sense –the sense of *P*. It is this last type of situation that is most relevant to our subject matter since this is where opacity, or non-extensionality raises its head.

[5] Namely, that the "star" visible in the evening sky after sunset is the same as the "star" that is visible in the east before sunrise, that is, the planet Venus.

Non-extensionality of a context means the non-substitutability of coreferential terms in distinct sentences *salva veritate* (that is, by preserving of truth value). Opacity means that one cannot deduce from 'J believes that The Evening Star is majestic' the sentence 'J believes that The Morning Star is majestic', even though 'The Morning Star' and 'The Evening Star' are coreferential names (designating Venus). The standard explanation for this non-substitutability is that it is not specified whether J knows the identity 'The Evening Star = The Morning Star', and it is only if this condition of knowledge is satisfied that one could deductively derive the second sentence from the first one.[6] Operators that create such situations of non-substitutability are called 'intensional operators'; examples include 'believes that …' and 'desires that …', which are also called 'hyperintensional', meaning that non-substitutability holds even for pairs of synonyms (that is, necessarily coextensive terms) in such contexts.[7]

The closest we can get to an instance of opacity in games is the case of the decomposed PD game, independently put forward by Pruitt (1967) and Messick & McClintock (1968). Pruitt decomposed the original PD game in such a way that now players were presented with two choice strategies (actions) of their own, A and B, which bring about utility outcomes resulting from combining these actions with the neutral (that is, independent of the other's actions) events 'give me' and 'give him'; players were told that the opponent is playing the same game and that the ultimate payoff is the result of the interaction between the two players. This fact is reflected in the algebraic structure of the decomposed game, as the payoffs of 'give me' and 'give him' add up to the payoffs of the original PD game.

Pruitt's ancestor PD game is represented in the payoff matrix in figure 1.

[6] Though see Saul 1997 for an argument for non-substitutivity even in simple sentences. Also, I offer (2013: 68–9) counterexamples to both the necessity and the sufficiency of knowledge of identities for intensionality to occur.

[7] Synonyms are not only coextensive but necessarily coextensive, that is, they will have the same extension in all possible worlds; yet, the belief operator, or the desire operator is hyperintensional in that one cannot guarantee substitution *salva veritate* in such a context since the subject might even fail to know that synonymy holds.

Fig. 1: Pruitt's ancestor PD game

		Player 2	
		A	B
Player 1	A	12, 12	0, 18
	B	18, 0	[6, 6]

As it is apparent, B is the dominant strategy for each player, hence (B, B), with outcome (6, 6) is a dominant strategy equilibrium. This, of course, is Pareto-suboptimal; it would have been better if the players "cooperate" and reach the (A, A) profile. This is the tragedy of the PD game. Players are doomed to end up in a suboptimal equilibrium. If only they could force one another to stay off-equilibrium and reach the Pareto-optimal outcome!

However, experiments with decomposed games point to a different story. Players seem to sometimes cooperate and sometimes defect, depending on which decomposition of the game they are dealing with. Thus, the decompositions of the ancestor PD game below generate different cooperative behavior in subjects. Matrix B elicits cooperation in about 20% of the subjects, whereas matrix C elicits it in 80% of them:

Fig. 2: Decomposition B

		Player 2	
		Give me	Give him
Player 1	A	6	6
	B	12	-6

Fig. 3: Decomposition C

		Player 2	
		Give me	Give him
Player 1	A	0	12
	B	6	0

It is easy to check that the payoffs for "me" and "him" add up to those in the ancestor game in Fig. 1, thus that game can be "reconstituted" from either B or C (and from countably many alternative decompositions).[8]

Now, this multiplicity of decompositions versus the uniqueness of the reconstitution does look like the most plausible analogue of Frege's idea of the multiplicity of senses versus sameness of reference. Consequently, I postulate that

> (*Reference of a game*) The reference of a game G, ρ(G), is its algebraic payoff structure.

> (*Sense of a game*) A sense of a game G, σ(G), is a mode of presentation of its algebraic payoff structure.

Furthermore, since in Frege sense determines reference (even though, as we have seen, reference does not determine sense):

> (*Reference determination in games*) For all games, G_1 and G_2, if they have identical sense, they have the same reference, that is, $\sigma_1(G_1) = \sigma_2(G_2) \supset \rho(G_1) = \rho(G_2)$

[8] For instance, in decomposition B, 'give me 12 and give him -6' played by both equals to 6 for each (i.e., the (B, B) profile in the original game):12-6 for each player since the game is symmetric. Similarly, in decomposition C, 'give me 6 and give him 0' played by both players equals to 6, and so on.

Finally, from the postulates above it follows that

(*Game identity*) For all games, G_1 and G_2, $G_1 = G_2$ iff $\rho(G_1) = \rho(G_2)$

(*Game identity*) is important in that, if found intuitively true, it leads to a normative claim to the effect that when reinterpreting a game, based on intensionality and possible senses, one should not change its payoff structure, that is, its reference, as it is tantamount to changing the game, hence, not a reinterpretation.

All this looks reasonable, but the problem is that we don't yet have a grasp on what could be meant by the *sense of a game*, that is, by 'mode of presentation of the payoff structure'. The problem is that a game is more complex than a sentence. Frege developed his theory for sentences of various degrees of complexity, but a game seems to be in a league of its own in that we have more variables that could be relevant to the idea of a mode of presentation of a game. For this reason, I think we should be liberal about assigning senses to games and allow for any of these variables to serve as the construction base for game senses.[9] Here are the variables I have in mind:

- Alternative *propositions* expressed by players' possible strategies (actions)
- Alternative *partitions* of the strategy space available to players.
- Alternative *identities* of the opponent as perceived by the player.

[9] This list might not even be exhaustive. One could also include, for instance, manipulations at the level of temporal order (simultaneous versus sequential structure) or at the level of information (perfect versus imperfect information versions of the same payoff structure. An example of the latter is the Manipulated Nash equilibrium, which is a refinement of subgame perfect equilibrium, propounded by Amershi, Sadanand, and Sadanand (1988). The idea is to start with a dynamic game of imperfect information, from which to generate a dynamic game of perfect information by simply deleting one or more information sets. The procedure by which equilibrium is now selected is called "forward induction", as opposed to the more well-known backward induction used in establishing subgame perfect equilibria in dynamic games of imperfect information. A profile is a Manipulated Nash equilibrium iff it is a Nash equilibrium of the parent game and it is subgame perfect in the new game. It is beyond the scope of this paper to analyze if and when manipulations of this type would still count as playing the same game, but it is something that is worth exploring.

Let me interject at this point, and note that although I do think that the sense of a game (which will, a fortiori, reveal the game's reference as well) will define that (type of) game, because such a liberal view of what is relevant to the notion of sense will entail that a game is more than the mere payoff structure (it includes possible actions or all players, for instance, as well as ways to structure that strategy space, such as in normal form or strategic form, etc.), I'm not committed to the claim that Pruitt's decompositions are merely senses of the PD game rather entirely different games. I have used Pruitt's decompositions to try to find an analogue of the Fregean notions, that is, *assuming* Pruitt is right to consider them as versions of the PD, we get a grip on the idea of a Fregean sense of a game. Given my criticism of Schick's ideas below, one could think I should rather be inclined to consider Pruitt's decompositions as ultimately illegitimate moves tantamount to changing the subject (the game).[10] The issue, therefore, hinges on whether these decompositions are legitimate ways of representing a game *in general*. The issue is not whether, once we accept them as legitimate representations of a game in general, they are representations of *the same game*; they clearly are.

In the remainder of the paper all these ways—propositions, partitions, and identities—in which a game could be non-extensional, i.e., reinterpretable/re-describable, will show up in various situations.

2. Schick on ambiguity in the PD

In his book, *Ambiguity and Logic,* Frederic Schick offers a new way to look at the infamous PD game. His idea (2003, 21–36) is that PD is a dilemma insofar as one is not prepared to formulate it in alternative ways, namely, to appeal to alternative ways for players to partition the contingency field of their opponents, i.e., the set of strategies available to them. If there are such acceptable ways of partitioning it under which there is no dominant strategy equilibrium that involves a Pareto-suboptimal outcome, then it is not true that in a PD type situation, necessarily, players fail to cooperate, if rational. They *could* cooperate and reach the Pareto-optimal state.

Let us again represent the PD game by way of the following payoff matrix.

[10] Many thanks to an anonymous referee for asking me to clarify this.

Fig. 4: Generic PD game

		Player 2	
		C	D
Player 1	C	3, 3	1, 4
	D	4, 1	[2, 2]

We have two players, each having two available strategies, C and D, which stand for cooperation and defection, respectively. We will call the set of strategies available to a player the player's *strategy space*. The possible payoffs of their interaction are given in the cells of the matrix. A cell of the matrix is called a *strategy profile*, while a pair of payoffs in a cell an *outcome*. Hence, the game has four possible outcomes, each corresponding to one of the four strategy profiles. Payoff pairs (a, b) represent payoffs for (Player 1, Player 2). A Nash equilibrium is a strategy profile such that none of the players has an incentive to move out of it, that is, to switch to a strategy other than what they have chosen.

The dilemma is thought to be that D is a dominant strategy for each player, but the dominant strategy Nash equilibrium of the game, (D, D), is Pareto-suboptimal. Each will choose to defect, because it is individually rational to do so, but the outcome is not the best one, since if both had chosen to cooperate, they would have got higher payoffs. So, it is a kind of collective rationality failure caused by following the rules of individual rationality.

Schick argues that the dilemma will disappear once we recognize that there are alternative ways the opponents can represent the situation to themselves, such that on those alternatives following one's equilibrium strategy does not lead to Pareto-suboptimality. One such alternative is a way to partition the opponent's strategy space, different from the one in Figure 4. Instead of representing his opponent's possible strategies as C and D, Player 1

could represent them as A, standing for 'I do the same as my opponent does', and O, standing for 'I do otherwise than as my opponent does'. If the opponent's set of strategies consists of A and O rather than C and D, then we get a new payoff matrix, represented in figure 5, Schick claims.

Fig. 5: Schick's reinterpreted PD

		Player 2	
		A	O
	C	[3,3]	1,4
Player 1			
	D	?[2,2]?	4,1

According to Schick, dominance in the matrix represented in figure 5 does not tell the players what to do. We have, he claims, two pure strategy Nash equilibria – (C, A) and (D, A)[11] – and, consequently, no dominant strategy equilibrium. What the new matrix tells us is that players *might* cooperate. What seemed to be a dilemma is now gone, if the new way of partitioning the strategy space is plausible.

Before expounding my core criticism of Schick's approach to the PD, let me put forward a couple of general comments.

[11] The reader will immediately notice that (D, A) does not appear to be a Nash equilibrium at all since Player 1 has an incentive to switch to C, according to this matrix representation. This is the reason I have flanked this strategy profile in the matrix by question marks. The problem is that, contrary to what Schick seems to think, the new matrix does not merely shift the payoff pairs around, but it distorts the very way we evaluate profiles for Nash equilibria. In this particular case, if Schick's point that cooperation is possible with the new representation is to be taken for granted, the only way we could make sense of (2, 2) being a Nash equilibrium is by equivocating on the expression "keep the opponent's strategy fixed", given that "fixed" will be conceptually linked to Schick's new strategy descriptions "doing the same as my opponent" and "doing the opposite of what my opponent does", which are non-rigid, hence, make the equivocation possible. See below for my criticism of Schick's move.

First, the PD is not really a problem *for game theory*. In fact, the PD is one of the clearest interactive decision problems, with a most straightforward solution: the dominant strategy Nash equilibrium (D, D). It is thought to be a dilemma for considerations lying outside the realm of what game theory is supposed to be a theory of, namely, a theory with explanatory, predictive, and normative power when it comes to human economic behavior[12].

Second, Schick's "solution" to the PD is, in fact, a problem for game theory: multiple Nash equilibria. The main reason game theorists have come up with a plethora of refined equilibrium concepts (called "Nash refinements" or "solution concepts") – such as subgame perfect equilibrium (Selten 1965), sequential equilibrium (Kreps and Wilson 1982), trembling-hand equilibrium (Selten 1975), and so on – is precisely to rule out or reduce the number of multiple equilibria in games. Viewed from this perspective, Schick's solution does more trouble than good.

Now back to the A/O partitioning. What are A and O supposed to be? One might think they are strategies. Schick himself at some point seems to think of them as just *doings*, that is, actions simpliciter, independent of one's description of them. He responds to the objection that it may be the case that A and O are not exhaustive of the opponent's strategy space, because she could have two more available strategies – cooperating whatever the other does, and defecting whatever the other does – by pointing this out:

> "This would be right if Jack believed that A and O were options for Jill, that each was an action she thought up to her. If she now expected to know what Jack will do before she chose, responding in kind and responding conversely would indeed be options for her, and silence regardless and talking regardless would be options too. But he knows she doesn't expect

[12] Indeed, as Gordon Tullock (1967) rightly pointed out, it is not straightforward even when considered from a normative point of view that the PD is such a dilemma. There are plenty of situations when it is in the larger community's interest to keep the members of some of its sub-groups in a PD-like situation: "The competitive market is an example. All the sellers of a given commodity could gain from higher prices, but each individual seller could gain from undercutting if the others raised their prices. The situation is a gigantic Prisoner's Dilemma. The whole point of antitrust legislation is to keep the sellers in this Prisoner's Dilemma, and much effort is invested by businessmen in attempting to get out of it. Clearly, however, the Prisoner's Dilemma serves a social purpose here, and there are probably many areas where this form of social control would be valuable." (p. 230)

> to have this advance information, and that she therefore doesn't think that A and O are up to her. In his eyes, given what he knows, A and O are not options for Jill. They remain what she might *do*, and, as *doings*, they are fully exhaustive. She will do one or the other; their disjunction leaves nothing out." (2003: 25–26, emphases as in original)

But to consider them as doings does not seem right. If they were actions, then the following two assertions by the opponent should be equivalent:

(1) I did A when Player 1 did C and I did A when he did D.
(2) I did A when Player 1 did C and did the same thing when he did D.

These propositions are not equivalent, under the interpretation of 'the same thing' as 'the same action, doing, or strategy'. Their non-equivalence is explained by the ambiguity of A with respect to what action is performed by Player 2; namely, A picks out C or D, depending on what player 1 does. A second reason why they are not strategies is that they cannot be used for defining exhaustiveness of alternative strategies of Player 1. When we say that C and D are exhaustive for Player 1's strategy space, we say that provided the opponent does the same thing, i.e. does not change his strategy, Player 1 can do either C or D and nothing more. But keeping the opponent's "strategy", A or O, constant does not amount to keeping his strategy constant.

A and O are rather *strategy descriptions*; they are functions from the elements of the set of Player 1's available strategies to player 2's strategies. But why should Player 1 partition Player 2's contingency field by appeal to such functions? Whatever representation he appealed to, the constraint would be that the opponent be plausibly taken to represent her own available strategies as the other represents them as being. But Player 2 could represent her own available strategies by appeal to such functions only if either she preferred ambiguous representations of what her available strategies are, or she had some extra utility

or disutility attached to imitating or not imitating the other player, respectively. None of these are part of what the PD is supposed to represent.[13]

Moreover, representing strategy spaces by appeal to ambiguous descriptions is sometimes tantamount to changing the subject, that is, the game. And, usually, it is no solution to a problem raised by a game to change the game to one in which the same problem does not emerge. In what sense can the matrix in fig. 5 be considered to have been derived from the PD?[14] Consider an analogy with the following game.

Suppose we have a game against Nature. Nature can equiprobably be in two states: the state of there being water, and the state of there being twin water.[15] Both states are such that there is a watery stuff. It is determinate that water is H_2O and that twin water is XYZ. We have Player 1 playing against Nature, namely he is presented with a quantity of watery stuff and asked to say whether it is water or twin water. He has two available actions: to say that it is water and to say that it is twin water. The payoff matrix is the following:

[13] An anonymous referee was critical of my possibly uncharitable approach to Schick's analysis, more precisely, of my running notions like "doings", "strategy" interchangeably, while they might have different meanings for Schick. Another, related issue was that I failed to take seriously Schick's distinction (2003: 32 – 33) between what he calls "a plight" and a dilemma, and the corresponding "option matrix" versus "decision matrix" distinction. The plight is supposed to be the initial situation of what the PD matrix presents to a player before any further structuring of the contingency field. The dilemma is one such structuring of it and has a decision matrix associated with it. But it is not the only one such possible structuring. This is the core idea of Schick's. I am not in disagreement with him about this much. I only think that his way of applying this general idea to the PD is forced and artificial. Finally, the referee also pointed out that Schick does have a story (2003:35) to justify his A/O interpretation of the "Prisoner's Plight". I find Schick's defense of it unconvincing and not something that we should prefer to standard game theory. Of course, this is an "ideological" disagreement. I prefer *reforming* standard game theory, which means adding variables to it but without questioning some elementary assumptions, such as the egoistic homo economicus model of the decision-maker, whereas Schick himself explicitly asserts that "In arguing that Plights need not be Dilemmas, my purpose was to suggest that they sometimes aren't. This would mean that people sometimes aren't economic agents, that they don't always S/T, that they are not always playing games (of the game theory sort)." (2003: 35)

[14] *Cf*. Pruitt's decompositions discussed in my section 1.

[15] This idea is, of course, inspired by Hilary Putnam's seminal papers on Twin Earth (1973, 1975)

Fig 6.

		Player 1	
		Water	Twin water
Nature	Water	1	0
	Twin water	0	1

The matrix represents Player 1's possible payoffs. The game is simple: if Player 1 says there is water and he is correct, he gets one unit. Otherwise, he gets nothing. *Mutatis mutandis* for twin water.

There are two pure-strategy equilibria (Water, Water) and (Twin water, Twin water). This being so, there is only a mixed-strategy unique equilibrium, namely, Player 1 choosing to say Water and Twin water, respectively, with equal probability. These two facts – that there are two equiprobable states for Nature and that Player 1 should use a fair coin to decide what to say – fit very well; we find this pairing intuitive.

Now consider instead that Player 1 opts for using 'water' ambiguously, as standing for H_2O when Nature is in state H_2O and XYZ when Nature is in state XYZ. She also uses 'twin water' ambiguously, as standing for XYZ when Nature is in state H_2O and H_2O when Nature is in state XYZ. Now the matrix looks like in figure 7:

Fig 7.

		Player 1	
		Water	Twin water
Nature	Water	1	0
	Twin water	1	0

Now we have two pure strategy equilibria, but an interesting new feature: Player 1 has Water as his dominant strategy. In other words, we have the pairing of a 50-50 chance of Nature coming up Water and Twin water with Player 1 deciding to say 'water' with certainty. This pairing, as opposed to the previous one, is far from being intuitive.

What happened with Player 1 preferring to use the expressions ambiguously is that he changed the game. There is no reasonable way now to think that the game in Fig. 7 is a way of representing the game in Fig. 6. That game was one of getting payoff 1 for correctly guessing the state of Nature, and nothing otherwise. This one is one of getting paid one unit for uttering the same expression whatever the state of Nature. Admittedly, in this latter game it is a bit easier to earn some money.

Let us then disambiguate A and O, such as to have the players' strategy spaces contain determinate actions. Of course, one way to do this is to appeal to C and D, but this is not an acceptable solution, as we are looking for *alternatives* to the original. Perhaps C/D is not the only solution. To this extent Schick's point is correct: we do have alternative ways of partitioning the opponent's strategy space, so what the PD represents as far as strategy spaces are concerned is open to interpretation. My point is that we should prefer alternatives that are not themselves built on ambiguities.

One way to proceed is to build the players' beliefs about what the opponent will do into their descriptions of actions A and O. So, we will have A{C} standing for 'I do the same thing as my opponent does and I believe that he does C'. *Mutatis mutandis* for A{D}, O{C}, and O{D}. Now, rather than having ambiguous action descriptions, we have unambiguous but co-designative pairs of distinct descriptions, i.e., one and the same strategy described in two different ways. Defection is referred to by both A{D} and O{C}; cooperation by both A{C} and O{D}. The new situation can be represented as in figure 8:

Fig. 8: Our reinterpretation of PD

		Player 2			
		A{C}	A{D}	O{C}	O{D}
Player 1	C	3, 3	1, 4	1, 4	3, 3
	D	4, 1	2, 2	2,2	4, 1

Of course, what we have on Player 2's side is not a partitioning of *strategies,* because there are two strategies, C and D, and they are exhaustive of the strategy space. But there can nevertheless be such a fourfold partitioning of *strategy descriptions*. Since there are only two possible beliefs that ground the belief-based descriptions of actions, namely, believing that the opponent does C versus believing that he does D, the four action descriptions are exhaustive of the action description space.

In the game represented in Fig. 8 we have two pure-strategy Nash equilibria: (D, A{D}) and (D, O{C}). What new information does this offer about the PD? Well, on the one hand, it just repeats the fact that (D, D) is a unique dominant strategy equilibrium, since the two equilibria are descriptions of the same outcome, namely (2, 2). In this sense, we are back to the original dilemma. But, on the other hand, this situation of multiple equilibria at the descriptive level – let us call it a situation of *multiple intensional Nash equilibria* – does tell us a couple of interesting things.[16]

[16] A referee complains that this is not the same game as the PD because the payoff structure is clearly different, namely, it is not symmetric, and the number of available actions is different. As the referee himself/herself observes, symmetry is only missing because I did npt represent the game both from Player 1 and Player 2's point of view. If we represent both, we get a four-by-four matrix. But that is also problematic, because that gives us 16 outcome profiles, not 4 as in the PD. But there is no real issue here, as far as I am concerned: the payoff structure refers to how much money you will and up with in each combination of strategies, and that is really the same here as in the original PD. The difference, as I say above, is at the level of descriptions of strategies, with no real effect on numbers and payoffs. The difference in strategies is illusory; it is a pseudo-difference. This might mean that it was

First, when being introduced to the PD game, one might be tempted to formulate the situation in terms of free riding behavior, that is, to equate following one's equilibrium strategy in the PD with being motivated by the incentive to free ride. What our new game representation shows is that this is not entirely correct. It says that one should be indifferent between (D, A{D}) and (D, O{C}), and so between A{D} and O{C}. A{D} stands for defection based on the belief that the opponent will defect. It is not a free riding incentive-based defection, but a fear-based defection. What the two intensional Nash equilibria say is that following one's equilibrium strategy in the PD is the same as mixing equally between free riding and fear-based considerations. As an empirical hypothesis, this could be confirmed if we conducted a poll on a large population of defectors and either found that roughly half of them defected because they wanted to free ride on others' cooperating and half of them acted so because they feared that the others will free ride, or that all of them had both kinds of motivation, equally weighted, when acting as they did.[17]

Second, one might be tempted to see and formulate the badness of the situation in terms of players' regretting their defection. For instance, Schick seems to do the same:

> "So if Jack and Jill are both rational, both of them will talk. But since each prefers the outcome of S, S (both of them silent) to that of T, T, they will both be sorry." (2003: 21)

Our multiple intensional Nash equilibria indicate that this is also wrong, if by 'both being sorry' one means both being sorry about what they do, namely, defection. The strategy profile (D, A{D}), that is, when one of the players defects motivated by fearing the other's defection, while the other actually defects, involves no regret: it is pretty straightforward that in this case the player's fears have been confirmed, so he should not regret his actions, but rather be satisfied that he had the right belief about what kind of player his opponent is, namely, a defector. But our multiple intensional Nash equilibria indicate that the player is or should be indifferent between this strategy profile and (D, O{C}). So, he shouldn't regret that

---

not even worth dealing with the task of redescribing strategies, but see below for why it is still useful and how we learn certain things even from these "fake differences" in strategies.

[17] There is, of course, a literature on this in experimental psychology. See, for instance, Hilbig et al 2018.

outcome either. It is quite intuitive to think that what our new game reveals is correct: defecting on the assumption that the other will cooperate means hoping that you will be able to free ride; if the other defects you shouldn't regret either your action, or your hope, because the action brings you more utility than the alternative action, while your hope ensures that you choose this particular action rather than the alternative. All potential regret on each player's end, according to our new model, is directed at the *opponent's* actions.[18]

However, alternative strategy space partitions at the descriptive level and our intensional Nash equilibria might give us new insights into other games, where standard game theory predicts the intuitively unlikely result.

3. Framing and the Fregean approach

I would like to clarify a few issues in this section regarding the way I see my proposal of a Fregean framework for games, or more exactly as an interpretation of intensional game theory, among extant approaches to framing that do not necessarily consider it as merely bias and irrationality.[19]

Let me start by pointing out that I do not see my proposal as a *rival* to extant theories of framing, but rather as a plausible way an analytic philosopher, as compared to an economist and a psychologist, could make sense of such framing phenomena. In this sense, what I put forward might well be thought of as complementary to current conceptualizations and explanations of framing. We, philosophers, are familiar with the Fregean framework of semantic analysis since it is ubiquitous in all our subdisciplines.[20] The idea of this essay is that we can even extend the Fregean semantic framework to *games* and it can serve as an umbrella framework, that is, as a generic scheme subsuming all theories that appeal to intensionality instead of bias.

---

[18] That the PD involves both fear and greed as motivations has been established in a seminal paper by Clyde Coombs (1973).

[19] I'm indebted to an anonymous reviewer as well as to Simon Columbus for pushing me to address these issues and suggesting very useful resources in this respect.

[20] I intend "Fregean" here in the most general fashion, as referring to any theory in philosophy which assumes, at the linguistic level, two dimensions of meaning: reference and sense. There is a stricter use of the adjective "Fregean", which refers to theories that adopt the "sense determines reference' thesis of Frege's in some form or other.

This brings me to the potential issue economists or psychologists might have with what they would dub as “merely linguistic” features of strategic interaction, such as the narrative that is used to describe or introduce the game, as opposed to the “real, mathematical structure” of it, that is, the matrix. Since analytic philosophy is really philosophy of language, sometimes called “linguistic turn” in philosophy, it is no surprise that my account here assumes language as central. This goes back to Frege, whose radical doctrine of *logicism* would have it that logic (that is, the formal language first-order calculus) is fundamental and mathematics is reducible to it. Regardless of whether one subscribes to this doctrine (or more recent versions of it) or not, we do think that linguistic analysis is the basis of understanding, including words, sentences, situations, worlds, and, I would add, games. Now, this does not mean that *any* feature of language is equally important or even relevant. Frege himself rejected any feature of language other than sense and reference as being relevant for meaning, hence, for understanding and communication. Indeed, his anti-psychologism—the doctrine that psychology is irrelevant to mathematics, logic, and philosophical understanding, and, strangely enough, even to psychology sometimes (*cf* Jacquette 2019: 351)—seeps through even in “Sense and reference”, where he strictly rejects the view that sense would have anything to do with psychology; psychology is about ideas, which are subjective, whereas sense is objective, hence, it is not the same as the notion of an idea. At the level of language, in effect, Frege considers that anything that is not either reference or sense, which are objective, are epiphenomena, not relevant for semantics (examples include, tone, poetic coloring, connotations of language use and related images, agent-relative associations (Jacquette 2019: ch 8).

Returning to games and the issue of the “real” meaning versus “merely linguistic” factors involved on game descriptions and given what I have just said in the previous paragraph, a Fregean is very comfortable with such a distinction. Indeed, one could see some level of similarity between the Fregean two-dimensional semantics that I am applying to games and some previous accounts which distinguish between "deep" and "surface" structure of games (Wagenaar et al., 1988) or "explicit" and "implicit" features (Hagen & Hammerstein, 2006), as well as the alternative accounts of what a game is, discussed by Ariel Rubinstein (1991), namely, the game as a physical situation versus the game as the

perception of a situation by agents.[21] If the narrative used to describe or introduce a game to the reader or to players is, intuitively, not part of the meaning of the game, then the Fregean has no issue with shoving it aside as not being part of the sense of the game.

A second issue I want to discuss is how this Fregean framework fits with extant theories of frames, especially from experimental psychology. In the context of experimental psychology of game playing, one classification divides frames into three types: focal points, valence frames (i.e., framing in terms of gains/losses), and context frames (i.e., changes to the name of the game, labels applied to choices, etc.) (Gerlach et al. 2017). The decomposed games I used earlier to introduce the sense/reference distinction as applied to games would arguably constitute a fourth category. It goes beyond the scope of this paper to discuss each of these in terms of how congenial a Fregean framework would be to them, so I will only say a few words about the one category that appears to me as probably the least fit for such a Fregean approach: context frames, that is, cases in which mere renaming of a game generate robust and stark changes in strategic behavior. There are many ways a game, or some component of it, in an experimental setting could be named. The player in a social dilemma type game, for instance, might be addressed as "you and the other", or "you and your opponent", or "you and your partner", etc. These might make a difference in a systematic way to how players behave. For example, describing the Prisoner's Dilemma as a "community" versus a "stock exchange" game induced stark differences in cooperation rates (Liberman, Samuels, and Ross's 2004).

It does seem as tough context frames are something a Fregean would reject as a candidate for game understanding and intensionality and rather include them in the category of linguistic but *extra-semantic* epiphenomena, hence, more fit to be judged as mere bias of rationality. The reason being that what a mere change of labels could elicit is at most something at the superficial and subjective level of associations, images, connotations. But I don't think this verdict is obvious. The whole point of the research on this type of frames is to show that even these apparently superficial, trivial-seeming changes lead to systematic and

[21] Though let me note that the similarity goes as far as the recognition of a descriptive *duality* in games; the difference is that unlike the Fregean framework put forward here, according to which both levels (reference/payoff and sense/frame) are core components of a game, the extant accounts of the above-mentioned descriptive duality tend to dismiss the 'surface structure' and assume extensionality when it comes to normative aspects of strategic behavior (e.g. Wagenaar et al. 1988)

robustly predictable behaviors, which means that the mechanism that creates the associations probably also has a deeper level which accounts for common game understanding or game representation among players. What exactly is involved in this mechanism is a matter of debate. There are few theories. One such theory interprets context frames in games of social cooperation as acting on the beliefs about the behavior of the other players and, as a result, inducing the conditionally cooperative player to change their behavior (Fischbacher and Gächter 2010). An alternative theory appeals to preferences about others' outcomes (social preferences) as the explanans of the framing effects (Ellingsen *et al* 2012). A third approach appeals to the idea of situation perception (which looks congenial in spirit to the Fregean idea that I am advertising here) and hypothesizes that players first interpret a certain game situation (for example, whether there are conflicting or corresponding interests and whether there are interdependencies) and then they play the game that results from such an interpretation (Columbus, Munich, and Gerpott 2020).

To reiterate a point I made earlier, I do not put forward the Fregean framework as a rival to these theories; if I did, it would have to be a psychological theory, just like the ones just mentioned, but it is not. It is a simply a recognition of the relevance of Frege style accounts of intensionality, without digging deeper into the exact psychological mechanism that is responsible for differences in behavior in framed games. The basic idea is that game representation goes beyond payoff structure, and it includes ways in which the *same game* can be seen/understood/conceptualized. This is in contrast with what Michael Bacharach points out as the two main approaches to explaining (or rather, explaining away) violations of classical game theory in games of social cooperation: *respecification* and *bounded rationality*:

> "Respecification theories explain behaviour usually thought of as an A-choice in a Hi-Lo by saying that the chooser is not in fact playing Hi-Lo but some related game G in which game theory does predict A. Bounded rationality theories explain A-choices in terms of limits on or lapses in rationality. In the Prisoner's Dilemma literature we find respecification theories in which G is, for example, an indefinite repetition of the Prisoner's Dilemma, or a Prisoner's Dilemma with transformed payoffs, and bounded rationality theories in which players have

limited depth or use magical reasoning" (Bacharach 2006: 47)

The Hi-Lo game type that Bacharach talks about, which is also the paradigm exemplar of a paradox of social cooperation that motivates the Bacharach/Sugden type intensional game theory (see footnote 3 above for references), is a game with multiple Nash equilibria, only one of which is collectively Pareto-optimal, like in this 2-by-2 matrix:

[Hi-Lo]:

		Player 2: **A**	**B**
Player 1:	**A**	5, 5	0, 0
	B	0,0	1, 1

The paradox of this game is that it is obvious that A is the rational choice of both players, yet classical game theory does not predict (A, A) as the unique equilibrium. Instead of "respecifying" the game (that is, changing it) or assigning less than full rationality to players (e.g. relaxing symmetry of information, or common knowledge, or some other such assumption about players), Bacharach appeals to an intensional approach in which players do not change anything about the payoff structure of the game, yet arrive to the cooperative strategy profile via a certain interpretation of themselves and their available actions within the Hi-Lo game, namely, as the part needed from them in order for the team to reach the collectively Pareto-optimal profile. The player thinks in terms of profile selection rather than in terms of best-response strategies as in classical game theory; also, players think of themselves as members of a group, the corresponding empirical hypothesis being that it is precisely Hi-Lo games that encourage such group identity or identification.

This approach to cooperation towards collective Pareto-optimality despite multiple equilibria appears to be focused on our third component of a game, player identity, that can be understood differently, hence can account for intensionality. If the player fails to identify with the group, then we have the behavior predicted by standard game theory, namely, mixing over the two strategies, A and B.

4. Intensional Nash equilibrium selection in other games

I will now define an Intensional Nash Equilibrium. A strategy profile is a vector having as elements the decisions taken by all players on an (imaginary or not) occasion of playing the game: $s = (s_1, \ldots, s_n)$. We denote by s_i player i's strategy and by s_{-i} the vector representing all the other players' strategies except i's. We denote by s^*_i player i's best response to s_{-i}; that is, the strategy that maximizes i's payoff, given the strategies played by all other players, which we denote by $\pi_i(s^*_i , s_{-i})$. A *Nash Equilibrium* is, then, a strategy profile, s^*, such that none of the players has incentive to deviate from her best response strategy, *conditional on none of the other players deviating from theirs*:

(*Nash Eq*) $s^*$is a Nash Equilibrium iff $\forall i \forall s'_i \, \pi_i(s^*_i , s_{-i}) \geq \pi_i(s'_i , s_{-i})$

An *Intensional Nash Equilibrium* is simply a Nash equilibrium under a certain *description*, δ, of whatever variable is relevant for a certain game under scrutiny –strategy, partitioning, or player identity.

(*Intensional Nash Eq*) s is an Intensional Nash Equilibrium iff $\exists \delta \; s^\delta$ is a Nash Equilibrium

Consider coordination games of the sort analyzed by Thomas Schelling (1960) with multiple equilibria, one of which has as its outcome a lower payoff than the others, but it is selected because of being the salient one. The outcomes of such equilibria are called 'focal points', or 'Schelling points'. Schelling point equilibrium selection takes us outside standard game theory, because according to the latter, the payoff structure tells us only that players might coordinate on any of the Nash equilibria, and so decide by randomizing over the available strategies. Yet, the idea that players will select the Schelling point equilibrium is empirically more adequate.[22] Consequently, Nash equilibria and Schelling point equilibria are concepts belonging to separate domains –standard, idealized game theory and behavioral

[22] See Mehta, Starmer, and Sugden 1994 for an experimental confirmation of focal point selection in Schelling style coordination games.

game theory, respectively. In other words, Schelling point equilibria are not normally considered Nash refinements.

Let us describe and represent such a Schelling game, call it ***G***. Suppose there are two players, 1 and 2, and each of them has three available strategies: to go to room 2215 of the skyscraper (A), to go to room 6218 of the skyscraper (B), and to go to the cafeteria situated nearby the skyscraper (C). Their objective is to meet each other to exchange some important information, and if they do not coordinate on the same place, they cannot re-coordinate afterwards; they lose the opportunity to exchange the information forever (so it is a one-shot game). However, as it happens, they don't like the cafeteria as much as they like any of the two rooms within the building where they could meet. Let us represent the game as in figure 9.

Fig. 9

G:

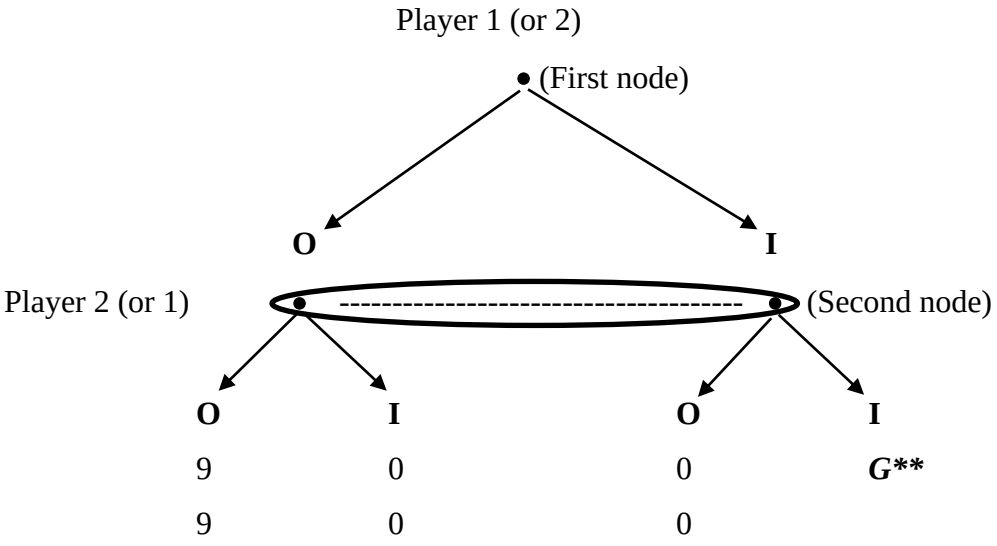
		Player 2		
		A	B	C
Player 1	A	[10, 10]	0, 0	0, 0
	B	0, 0	[10, 10]	0, 0
	C	0, 0	0, 0	[9, 9]

The game has three pure-strategy Nash equilibria – (A, A), (B, B), and (C, C) – and a mixed-strategy equilibrium in which each player randomizes with equal probability over the three strategies. However, Schelling's insight was that such a game has nevertheless a unique empirically selected Nash equilibrium, namely (9, 9). This outcome has something that distinguishes it from the rest of the equilibrium outcomes; it breaks the 10 – 0 symmetry of the payoff representation.

As I have mentioned before, such equilibrium selection has no place in standard game theory. But our intensional Nash equilibria may explain such an equilibrium selection *by way of considerations pertaining to standard game theory*. Instead of players describing the available strategies by A, B, and C, they can, and it is empirically plausible that they will, describe them as I: 'meeting inside the skyscraper' and O: 'meeting outside the skyscraper'. The fact that such a partitioning of descriptions is available is, of course, determined by the fact that the two rooms resemble each other more, from the point of view of what is described in the game, than each of them resembles the cafeteria. Besides this, the players could represent the game as if it were sequential, with a first and a second mover. This is unproblematic since the game is symmetric, there is common knowledge and there is no first (or second) mover advantage in the game. Hence, representing it *to themselves* as sequential is consistent with them knowing it is not sequential. The game, call it ***G****, now looks different, though the payoffs and the strategy profiles of the original game have not changed at all. Now Player 1 (or Player 2) decides whether to enter the building or not. Not entering the building is equivalent to going to the cafeteria. However, entering the building yields another game: players must coordinate on one of the two rooms. Call this subgame ***G*****; it will be played conditional on Player 1 (or Player 2) entering the building. We can represent the situation as in figure 10:

Fig. 10

G*:



Subgame ***G****** is represented below.

Fig. 11

G:**

		Player 2	
		A	B
Player 1	A	10, 10	0, 0
	B	0, 0	10, 10

Game **G**** is itself a coordination game, with two pure-strategy Nash equilibria – (A, A) and (B, B) - and a mixed-strategy equilibrium – (½ A, ½ B), and an expected payoff of 5. Given these facts, game **G*** has only one subgame perfect equilibrium, the one in which both players choose the cafeteria. This is because the expected utility of ending up in this strategy profile (O, O) is 9 and it is very easy to coordinate on it once the G** coordination game is seen as risky.

Note that sequential representation is necessary but insufficient for this unique equilibrium selection. It can be checked that if players had not appealed to redescribing the strategy space the way they did but had only appealed to a sequential representation of strategies A, B, and C, they would have had to choose between two equilibria – (A, (A, B, C)) and (B, (A, B, C)) – yielded by their both choosing A or B, respectively.

The unique equilibrium we have obtained is an example of what I have put forward and dubbed intensional Nash equilibrium, – a Nash equilibrium under a certain description of the strategy space. The fact that it is a unique intensional Nash equilibrium explains the fact that it is selected. So, Schelling point equilibrium selection has a *rational* explanation in the classical, orthodox sense of rationality. Of course, why the strategy space could be described the way it was is explained by a certain outcome having the property of being a Schelling point. Further, one could contemplate a change in our game so that the expected payoff from

playing subgame **G**** is equal to the payoff associated with (C, C), namely, 9. It can now be argued that players still choose C, yet there are, according to our model, two intensional Nash equilibria – (O, (I, O)) and (I, (I, O). This may be the case, but still a rational explanation is that players refuse to enter the building and play another game with risky outcomes. So, risk aversion may play an important role. Further, one may note that if we keep raising the expected payoff of **G****, we will clearly have two equilibria, while if players choose their Schelling strategy, we have again a mismatch between Nash equilibrium selection and Schelling point equilibrium selection. I reply that once we raise the expected payoff of **G**** to a sufficiently high value, outcome (9, 9) ceases to be a Schelling point in the sense of being *both focal and still attractive*. There is a lot of vagueness between it being and not being a Schelling point, but it is plausible that for an expected value of, say, $1000 – for which we need outcomes of ($2000, $2000) – (C, C) is not a Schelling point any longer. Strategy A and B can now be described as 'the strategies that can get me a lot of money', as opposed to C, which becomes 'the strategy that does not get me any significant gain'. Rather than having a point, we now have what might be called a *Schelling region* – ((A, A), (B, B)). Hence, the idea of intensional Nash equilibria as explanatory of Schelling point equilibrium selection is not invalidated.

Viewed from this perspective, intensional Nash equilibria can even be considered as some sort of Nash refinement. More exactly, we can say that given a payoff structure, all the intensional Nash equilibria are also Nash equilibria, but not all the Nash equilibria are intensional Nash equilibria *on all descriptions of the strategy space*.

Finally, there are cases when a certain equilibrium outcome has some property that makes it the uniquely eligible outcome to be referred to by a catchy description, and so become a unique intensional Nash equilibrium. Such a case is I think the fair division equilibrium outcome of the Bargaining Problem, whose unique selection was axiomatized by John Nash (1950). The problem consists of two rational players having the task of dividing a resource, say, 100 dollars, between themselves. Suppose the smallest non-zero amount that one can get is one dollar. Then we have 101 Nash equilibria, each of them corresponding to a Pareto-optimal outcome. However, there is one outcome such that we have a catchy name for it – the fifty-fifty division. The name that we have for it is 'the fair division'. All other outcomes are such that the strategies whose Cartesian product corresponds to them will receive different descriptions from the two players involved in the situation. For instance, the

70/30 outcome is the product of two strategies such that one can be described as favorable to one of the players, the other as unfavorable to the other player. Adopting a terminology of David Lewis's (1983: 371), the fifty-fifty outcome is *eligible* to be uniquely described the way we describe it; or, adopting a nowadays-fashionable way to express Lewis's point, the outcome is a *reference magnet* – it has some property that attracts a unique way to refer to it. Given this intrinsic magnetism of the outcome, players will have a uniquely denoting description of it, such that the outcome will be the *unique* intensional Nash equilibrium under that description.[23]

5. Identity, rationality, and intensionality

The core idea of intensional decision and game theory is that violations of extensionality in decision problems is not always tantamount to irrational behavior. The details of how this is supposed to work are being developed. It is an open project, far from completion, and for that reason it is exciting. In this essay, the metatheoretical guiding principle that I have assumed was that when speculating about ways to reinterpret games in the spirit of non-extensionality one should be as conservative as possible, that is, stay close to classical, *homo economicus* based expected utility maximization. Among other corollaries that might follow, the most important one is that one should therefore not change the payoff structure of games when speculating about intensionally non-equivalent versions of it. This, of course, reduces the range of possible maneuvers, yet it is a challenge that might elicit creative moves and in a way that would make even defenders of orthodox game theory accept intensionality in games. There are several non-classical moves in the history of reforming classical decision and game theory, such as alternative utility functions[24] and evolutionary game theory[25], and

[23] I am not claiming here that the way the intensional Nash equilibrium is selected is different and somehow better than Schelling's idea of a focal point. One way, in effect, to identify a focal point in a situation like this is to observe that it has a description (such as "the fair division", or "the equal distribution"), unlike all the other equilibria. Thanks to a referee for asking me to clarify this.

[24] That is, utility functions that do not maximize utility based on the assumption of egoism. For example, the model of the Kantian agent, who maximizes the categorical imperative (Alger and Weibull 2003) or the inequity aversion model (Fehr and Schmidt 1999), according to which the agent minimizes inequality between agents' payoffs.

they are fruitful endeavors. But my main goal here was to promote the idea of non-extensionality in the tamest way, that is, in a way compatible with even the most conservative approaches.

The second thing I would like to point out is related to the concept of identity in the context of games. Earlier I mentioned the identity of the opponent in the player's perception as one of the variables that could be relevant for the intensional approach to games. Differential behavior based on *who* the players think the opponent is is not uncommon to observe in experiments.[26] By "identity", in this context, we don't mean numerical identity, but identity under a type-description, e.g. "an F-type player", which is just what the Fregean sense is supposed to contain (*cf.* Frege when he says that the sense is wherein the mode of presentation is contained). At the same time, when it comes to extensionality, the notion of identity that is relevant is the numerical one. Cases of intensionality (that is, violations of extensionality) are cases of numerical identity that coexists with non-identity of sense or mode of presentation. The problem of irrationality in connection with intensionality does not arise in the philosophy of language. There is nothing irrational (on the contrary!) and nothing blameworthy about Oedipus not having a desire to marry his own mother, yet simultaneously being delighted to marry Jocasta (Jocasta being, unknown to Oedipus, identical to his biological mother). But it does arise in decision theory, and here there is room for debate as

---

[25] Pioneered by John Maynard Smith (Maynard Smith and Price 1973, Maynard Smith 1982), evolutionary game theory is a radical departure from classical game theory in that agents are not decision makers in the standard sense of being able to deliberate and choose from elements of a choice set; rather they are genetically programmed to follow one strategy. The resulting equilibrium concept, called "evolutionarily stable strategy", is a population level variable corresponding to a steady state of the dynamic system emerging from the interactions of subpopulations.

[26] *Cf.* Sally Blount's (1995) experiments with the repeated ultimatum game, where players accept lower offers from the opponent if they think it is a computer they are playing against (a random number generator); conversely, they tend to punish human opponents with more ease, regardless of whether the opponent is presented as having a stake or not in the game. In the ultimatum game players have a chance to split a fixed sum of money, say 10 dollars; one of them offers the other a share, which the latter can accept or reject. If she rejects it, then none of them receives any money. If she accepts the offer, then each of them gets the corresponding sum according to the split. The game has a thousand Nash equilibria (1 cent/999 cents, 2 cents/ 998 cents, etc.). Standard game theory prescribes and predicts that the receiver should accept any offer. Experiments do not support this, except that when played against a computer, the rejection rate is close to zero. This difference has also been observed at neural level, in fMRI studies (Sanfey et al 2003, Civai 2012).

about cases in which the agent is blameworthy or not (and to what extent) for violations of extensionality. The intensionalist decision and game theorist's position is that the agent is *not always* blameworthy when violating extensionality. The questions then are: in what types of situation and to what extent in each type? A conservative position here could be that the agent is not guilty of irrationality only when the extensionality violation is due to excusable ignorance of an identity, where "excusable ignorance" would mean something like:

> (*Excusable ignorance*) S is excusably ignorant of a = b in a decision problem *D* iff
>
> (a) a = b is true
>
> (b) S does not believe that a = b
>
> (c) the cost to S to learn that a = b is higher than any of S's payoffs in *D*

A liberal position could be the one put forward in Bermudez 2018 (and developed in Bermudez 2020), under the name "ultraintensionality". Ultraintensional contexts are supposed to be contexts that permit rationality-preserving failures of substitutivity of *known identities*. Under this liberal interpretation, the agent S could behave differently in two distinct game representations (senses), $\sigma_1(G_1) \neq \sigma_2(G_2)$, with the same reference, even while understanding that $\rho(G_1) = \rho(G_2)$. Though this level of liberalism regarding rationality-preserving intensionality might seem exaggerated, the idea is that valuations are precisely the kind of situation in which we might behave this way. Schick's (1991) discussion of an example from George Orwell's autobiographical essay "Looking back on the Spanish Civil War" appears to be a convincing example. Orwell describes a situation in which during his time with the International Brigades fighting the Fascists in Spain he was faced with the choice of shooting or not shooting an unarmed, half-dressed man who jumped out of the trenches, holding up his trousers with both his hands while running. Orwell's decision problem is whether to shoot that person or not, and he is inclined not to shoot him under the mode of presentation "half-dressed man carrying his trousers in his hands" even though he does have a preference to shoot Fascists, hence, to shoot the man under the mode of presentation "a Fascist".

Perhaps a middle position is also available, one that is conservative enough to please even old school normative game theorists while simultaneously making sense of

intensionality at the level of player-identity. Like with my earlier analysis of the Schelling type coordination games, I think the key here is not equilibrium existence, but equilibrium *selection*. Intensionality is not only a fun fact but it is useful in equilibrium selection situations when we have multiple equilibria. Behavioral economics and the psychology of decision-making, of course, have a simple and straightforward solution in the spirit of revealed preference: simply let many people play the game, observe the percentages for each Nash equilibrium, and then try to explain the psychology behind the leading equilibrium strategy, if any such shows up. The normative approach reverses the process: let us find various alternative modes of presentation of one or more aspects of the game (payoff representations, strategy space representations, strategy space description partitions, player identity representations) and see how they could differentially affect equilibrium selection. This does not mean (and it is not what I claim) that considering intensionally non-equivalent representations of the game will always or even typically reduce the number of Nash equilibria, but that sometimes it will, and even when it does not, it might well make one equilibrium stand out in some way and therefore likelier to be selected in experimental settings.

To accommodate this more traditional, normative approach, we could focus on games with multiple pure strategy Nash equilibria where the available actions can be connected to, or correlated with, a type of player (e.g. bold versus submissive in a *Hawk-Dove* discoordination game) and in which although type signaling does not occur, the mode of presentation under which the opponent appears to the other player is going to make a difference, intuitively, to equilibrium selection. Furthermore, unlike the conservative view mentioned above in guise of the principle of excusable ignorance of an identity, we do not think of the players as having an obligation, or to invest any resources whatsoever in trying to find out that the two modes of presentation are of the same opponent.

To exemplify, I will use the game known as *Battle of the Sexes*. Fig. 12 is a depiction of the standard, generic, textbook version of the game. The story is the following. We have a couple, man and woman, who prefer to spend the afternoon together rather than in solitude. They have two possible options of where to go: horse racing and opera. The man prefers the horse race to the opera, while the woman has the reverse preference ordering. Spending the afternoon alone brings no utility to anyone.

Fig. 12: ***Battle of the Sexes (generic)***

		Man	
		Racing	Opera
Woman	Racing	[1, 2]	0, 0
	Opera	0, 0	[2, 1]

As is apparent, the game is symmetric and there are two Nash equilibria: both the man and the woman going to the horse racing and both going to the opera. Now consider versions in figure 13a and in figure 13b of this game. Our story now is loosely based on Baroness Orczy's novel, *The Scarlett Pimpernel*. We have Marguerite Blakeney, the main female character from the novel, who is disillusioned by her husband, Sir Percy Blakeney's fap lifestyle and values, while being mesmerized by the daring exploits of the mysterious hero, The Scarlett Pimpernel, who saves aristocrats from the guillotine during The Reign of Terror in the early days of the First Republic in the French Revolution. Marguerite does not (yet) know that Sir Percy = Pimpernel. Finally, it is part of the story that Sir Percy is also not impressed by his wife any longer, ever since he heard that she was instrumental in having Marquis de St. Cyr and his entire family sent to the guillotine. In version **a** below, we depict Marguerite's perspective on the game when the mode of presentation of the "opponent" is that of Sir Percy, while version **b** depicts the man as Pimpernel.

Fig 13a: *Battle of the Sexes – Sir Percy* version

		Sir Percy	
		Racing	Opera
Marguerite	Racing	[1, 9]	0, 0
	Opera	0, 0	[9, 1]

Fig 13b: *Battle of the Sexes – Scarlet Pimpernel* version

		Pimpernel	
		Racing	Opera
Marguerite	Racing	{[199, 100]}	0, 0
	Opera	0, 0	[200, 1]

The payoffs reflect Marguerite's, Percy's/Pimpernel's attitudes to one another, namely how excited they are to be in one another's company. Marguerite and Percy/Pimpernel are not overly excited about one another's company (though they prefer it to solitude), whereas Marguerite is thrilled about Pimpernel's. Though we have not reduced the number of equilibria in 13b as compared to figure 13a, nevertheless the game "looks better" in 13b, in the sense that it would be a lot less crazy or less inexplicable[27] for an analyst to hypothesize that Marguerite and Pimpernel will choose racing and end up in the best of the two

[27] Unlike in the case of decomposed PD games that I presented in section 1, where it is puzzling why subjects behave so radically differently when confronted with various such decompositions.

equilibria[28]. Marguerite's reasoning would be that the real prize here is Pimpernel's proximity, hence it is not worth risking but rather just go for racing, which is what Pimpernel prefers. Pimpernel ls a lot more interested in racing and barely gets any pleasure from opera with Marguerite. So, it is intuitive that the (Racing, Racing) will be selected in games played this way. In version **a**, on the other hand, we have a perfectly symmetric double equilibrium, hence, the best we (or Marguerite and Percy) can do is flip a coin.

***

To conclude, this essay is an addition to the so far regrettably underrated non-extensional approach to decision theory and the philosophy of rationality. I have focused on interactive decision theory, i.e., on games, and within that on the noncooperative, one-shot type. The take-home message is that violations of extensionality are not (or not always) merely errors of reasoning, but the way we humans sometimes reason in strategic context that can present themselves under several modes of presentations. I have opted for programmatic conservativeness for the time being, in the form of trying to not deviate too much from classical normative game theory. However, as the reader might well intuit, there is probably even more potential for intensionality if we allow ourselves more audacious departures from this standard. In any case, what I have been trying to show is that the intensionalist approach to decision and game theory is something well worth exploring.[29]

References

Ahn, D. S., & Ergin, H. (2010). Framing contingencies. *Econometrica* 78(2): 655–695.

[28] A referee observes that the numerical values of the payoff have changed radically in the "Pimpernel" version and wonders why this I not a different game. I reply that the structure and the number and identity of the equilibria have not changed, that's why this is an interpretation of the original game, not a different one.

[29] Acknowledgments: I am grateful to the audience at Bilkent University, where I presented this paper as part of the Work in Progress Seminar. I am indebted to three referees (two anonymous ones and Simon Columbus) whose illuminating comments improved the paper. I am also grateful for the continued support of my research by TÜBITAK (The Scientific and Technological Research Council of Turkey).

Amershi, A. H., A. B. Sadanand, V. Sadanand (1988) Manipulated Nash Equilibria I: forward induction and thought process dynamics in extensive form, *Working Paper No. 928*, University of British Columbia, Vancouver.

Alger, I., and J. W. Weibull (2013) *Homo moralis*. Preference evolution under incomplete information and assortative matching, *Econometrica* 81 (6): 2269–2302.

Bacharach, M. (1999) Interactive team Reasoning: A contribution to the theory of cooperation'. *Research in Economics* 53,117–147.

Bacharach, M. (2001) Superagency: Beyond an Individualistic Theory of Games. In Johan van Benthem, ed., *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2001)*, San Francisco: Morgan Kaufmann. 333–337.

Bacharach, M. (2006) *Beyond Individual Choice. Teams and Frames in Game Theory*, Princeton University Press.

Bermudez, J. L. (2009) *Decision Theory and Rationality*, Oxford University Press.

Bermudez, J. L. (2018) Frames, rationality, and self-control. In J.L. Bermudez (ed.) *Self-control, Decision Theory, and Rationality*, Cambridge University Press.

Bermudez, J. L. (2020) *Frame it Again. New Tools for Rational Decision-making*, Cambridge University Press.

Blount S. (1995) When social outcomes aren't fair: The effect of causal attributions on preferences, *Organizational Behavior and Human Decision Processes* 63 (2): 131–144

Bourgeois-Gironde, S. and R. Giraud (2009). Framing effects as violations of extensionality. *Theory and Decision*. 67 (4): 385–404.

Civai, C. e al (2012) Equality versus self-interest in the brain: Differential roles of anterior insula and mfredial prefrontal cortex, *NeuroImage* 62 (1): 102–112.

Columbus, S., J. Münich, and F. H. Gerpott, (2020). Playing a different game: Situation perception mediates framing effects on cooperative behaviour. *Journal of Experimental Social Psychology* 90: 104006.

Coombs, C. H. (1973). A reparameterization of the prisoner's dilemma game. *Behavioral Science,* 18(6): 424–428.

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

Fischbacher, U. and S. Gächter, (2010) Social preferences, beliefs, and the dynamics of free riding in public goods experiments, *American Economic Review* 100(1): 541-556.

Frege, G. ([1892] 1948). Sense and reference. *The Philosophical Review* 57(3): 209–230.

Frisch, D. (1993) Reasons for framing effects. *Organizational Behavior and Human Decision Processes* 54(3): 399–429

Gerlach, P., Jaeger, B., and Hertwig, R. (2017). Cooperation needs interpretation—A meta-analysis on context frames in social dilemma games. In P. Gerlach, *The social framework of individual decisions: 570+1 experiments in (un)ethical behavior* (pp. 9-40). Doctoral dissertation, Humboldt-Universität zu Berlin, Berlin, Germany.

Gold, N. and Ch. List (2004) Framing as path dependence. *Economics and Philosophy* 20: 253– 277.

Guth, W., R. Schmittberger and B. Schwarze (1982) An experimental analysis of ultimatum bargaining, *Journal of Economic Behavior & Organization* 3 (4): 367–388.

Hagen, E. H., and Hammerstein, P. (2006). Game theory and human evolution: A critique of some recent interpretations of experimental games. *Theoretical population biology*, 69(3): 339-348.

Hilbig, B. E., P. J Kieslich, F. Henninger, I. Thielmann, and I. Zettler (2018). Lead us (not) into temptation: Testing the motivational mechanisms linking honesty-humility to cooperation. *European Journal of Personality* 32(2): 116–127.

Jacquette, D. (2019) *Frege*. *A Philosophical Biography*, Cambridge University Press.

Kreps, D. and R. Wilson (1982) Sequential equilibria, *Econometrica* 50 (4): 863–894.

Kühberger, A. (1998) The influence of framing on risky decisions: a meta-analysis. *Organizational behavior and human decision processes* 75(1): 23–55.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes* 76(1), 149–188.

Lewis, D. K. (1983) New work for a theory of universals. *Australasian Journal of Philosophy* 61: 343-378.

Maynard Smith, J. and G. R. Price (1973). The logic of animal conflict. *Nature*. 246 (5427): 15–18.uy

Maynard Smith, J. (1982) *Evolution and the Theory of Games*. Cambridge University Press.

Mehta, J., C. Starmer and R. Sugden. 1994. 'The Nature of Salience: An Experimental Investigation of Pure Coordination Games'. *American Economic Review* 84: 658–673.

Messick D. M. and McClintock C. G. (1968) Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology* 4(1): 1–25.

Nash, J. F., Jr. (1950) The bargaining problem. *Econometrica* 18: 155-162.

Pruitt D. G. (1970) Motivational processes in the decomposed Prisoner's dilemma game. *Journal of Personality and Social Psychology* 14(3): 227–238.

Putnam, H. (1973) Meaning and reference, *Journal of Philosophy* 70 (19): 699–711.

Putnam, H. (1975) The meaning of 'Meaning', *Language, Mind and Knowledge. Minnesota Studies in the Philosophy of Science, vol. 7*, Keith Gunderson (ed.) Minneapolis: University of Minnesota Press, 131–93.

Rubinstein, A. (1991) Comments on the interpretation of game theory, *Econometrica* 59(4): 909–924.

Sanfey, A.G. et al (2003) The neural basis of economic decision-making in the Ultimatum Game, *Science* 300 (5626): 1755–1758.

Saul, J. M. (1997) Substitution and simple Sentences. *Analysis* 57 (2): 102–108.

Schelling, Th. C. (1960) *The Strategy of Conflict*, Harvard University Press, Cambridge.

Schick F. (1991) *Understanding Action: An Essay on Reasons*. Cambridge: Cambridge University Press.

Schick, F. (1992) Allowing for understandings, *Journal of Philosophy* 89 (1): 30–41.

Schick, F. (1997) *Making Choices: A Recasting of Decision Theory*, Cambridge University Press.

Schick, F. (2003) *Ambiguity and Logic*, Cambridge University Press.

Selten, R. (1965) Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit. Zeitschrift für die gesamte Staatswissenschaft 121 (2): 301–324, 667–689.

Selten, R. (1975) A reexamination of the perfectness concept for equilibrium points in extensive games". *International Journal of Game Theory* 4 (1): 25–55.

Sugden, R. (1993) Thinking as a team: towards an explanation of nonselfish behavior'. *Social Philosophy and Policy* 10: 69–89.

Sugden, R. (2000) Team preferences. *Economics and Philosophy* 16: 175–204.

Sugden, R. (2003) The logic of team reasoning, *Philosophical Explorations* 6: 165–81.

Tullock, G. (1967) The Prisoner's Dilemma and mutual trust. *Ethics* 77 (3): 229-30.

Tversky, Amos; Kahneman, Daniel (1981) The Framing of decisions and the psychology of choice. *Science* 211 (4481): 453–58.

Wagenaar, W. A., Keren, G., and Lichtenstein, S. (1988). Islanders and hostages: Deep and surface structures of decision problems. *Acta Psychologica*, 67(2): 175–189.