Similarity as an Extension of Symmetry and its Application to Superrationality

Carlos Maximiliano Senci[1]

http://orcid.org/0000-0001-9131-3843
[1] *IIESS-Conicet and Departamento de Humanidades-UNS*
Bahía Blanca, Buenos Aires Argentina
maxsenci@gmail.com

Fernando Abel Tohmé[2]
https://orcid.org/0000-0003-2988-4519

[2] *IMABB-Conicet and Dept. of Economics-UNS*
Bahía Blanca , Buenos Aires
Argentina
ftohme@gmail.com

Article Info
CDD: 501
Received:05.11.2020; Revised: 01.02.2021; Accepted: 05.03.2021
https://doi.org/10.1590/0100-6045.2021.V44N2.CF

Keywords:
Game theory
Symmetry
Similarity
Superrationality

Abstract: In this paper we present a concept of *similarity* in games, on which to ground alternative solution concepts, some of which differ from the classical notions in the field.

In order to do this we impose a constraint on players' beliefs that amounts to a variant of the well-known symmetry principle in classical bargaining theory. We show how this similarity relation helps to identify different Nash equilibria in games, and how these "similar Nash equilibria" can be extended to non-symmetric games. While the notion is normative, it is nonetheless inspired by phenomena in which similarities between players lead to outcomes detected in behavioral studies. We study the strategic properties of the concept of similarity and discuss its relationships with Hofstadter' notion of superrationality.

1 Introduction

As we know, Game Theory studies situations of interaction between intentional agents. By 'interactive situations' we understand those in which more than one player intervenes, and in which the potential results are the product of the joint actions (simultaneous or not) of several players. Game theory proposes solution concepts, that is, rules indicating the choices to be made by the players in order to obtain results that satisfy some theoretical criteria of appropriateness. Solution concepts yield a set of strategy profiles (eventually the empty set) for each finite strategic game. In this sense, Game Theory is normative, since it prescribes solutions that disregard the actual decision-making processes carried out by flesh and blood humans. It rather provides solutions to be chosen by idealized agents.

Nash Equilibrium (from now on denoted NE; Nash, 1950; Myerson, 1991, p. 105) is the most important solution concept in Non-Cooperative Game Theory. Informally, a NE is a profile of strategies such that none of the players has incentives to change her strategy, given the strategies chosen by the other players. It is well known that the epistemic conditions required by the NE of the players are quite restrictive. In par-

ticular, in the simplest case of two players, the NE requires rationality and mutual knowledge of rationality, i.e. both players must be rational and know that they are rational (Aumann and Brandenburger, 1995). These players are clearly idealized agents.

In many ubiquitous situations the NE does not provide satisfactory solutions. The typical case is the Prisoner's Dilemma, but there are other well-studied games, such as the Stag Hunt or Hi-Lo (Colman and Gold, 2018), in which there are payoff-dominant results that the standard theory cannot justify. An outcome is dominant if all players get larger payoffs than with any other results.

In many games in which there are multiple NE, nothing indicates which one should be chosen. Then, it becomes necessary to arbitrate solutions that refine the concept of NE. The NE refinement program has generated a large literature (see, in particular, Harsanyi and Selten, 1988). The attempt to refine concepts of strategic rationality does not seem to have found a general solution like the one sought by philosophers, able to accommodate all the contingencies that arise in relation to the justification of any solution. In the same way, it does not seem that they have been able to accommodate in the vast array of anomalies that arise from the experimental results. Behavioral Game Theory can be seen as a body of solution concepts corresponding to those results, going beyond the limits of rational choice (Ross, 2014). But this proliferation of solution concepts, while an indication of the impetuous development of Behavioral Game Theory, also testifies to a certain failure, partial at least, to provide a unified framework (Tohmé et. al., 2017).

In this search of alternative solutions, sometimes it is convenient to assume that players are somewhat similar to each other. A similarity relation could refer

to normative aspects, such as rationality. But empirical aspects without normative content may be also relevant. Such is the case if we think about the abilities or qualities of the players. This leads to solutions based on the strategic depth of players (k-level reasoning; Camerer et al., 2004), their level of conformity to norms (Bernheim, 1994), or their levels of inequity aversion (Fehr and Schmidt, 1999), among others. The similarity between the players reduces the uncertainty experienced by each of them when facing players she considers similar to her. This behavioral phenomenon has been documented in a series of recent studies, in which different similarity primings had effects on coordination or cooperation (eg, Fischer, 2009; Di Guida and Devetag, 2013; Mussweiler and Ockenfels, 2013 ; Chierchia and Coricelli, 2015; Rubinstein and Salant, 2016).

There are different attempts to capture similarity relationships between players. In this paper we provide a formal representation of the concept of similarity, showing how it helps to identify different NE in games. We also show how these 'similar Nash equilibria' can be extended to non-symmetric games, and under additional restrictions we will show how to extend it to superrationality.

Superrationality was proposed by Hofstadter (1983, 1985) to account for the cooperative solution in the Prisoner's Dilemma (PD) (Flood and Dresher 1952). It is one of the alternative solutions that have been advanced according to which players could do better than the mutual defection in the PD. Proponents of the cooperative outcome usually enrich the description of the PD with an additional assumption according to which the players are somehow *similar*. Most notably, one of those lines of thought focuses on *symmetry* as the basic argument for cooperation, rooted in the alleged ability of the players to recognize that they are facing

the same situation and that their strategies are the same (Davis, 1977; Campbell, 1989; Rapoport, 1960).

Hofstadter devised the idea of *superrational thinkers* as a solution for symmetric dilemmatic situations, that is, one in which the incentives are the same for all the players. Superrational thinkers reason their way to the solution by applying a meta-rationality norm implying that every thinker knows what the rational answer is, and knows that every other player knows this fact. So, superrational thinking is not a simple property of an individual but a relational property:

> "You need to depend not just on their being rational, but on their depending on everyone else to be rational, and on their depending on everyone to depend on everyone to be rational-and so on. A group of reasoners in this relationship to each other I call superrational. Superrational thinkers, by recursive definition, include in their calculations the fact that they are in a group of superrational thinkers." (Hofstadter, 1985, p. 748).

Since superrational thinking applies to symmetric situations, the (unique) rational solution identified at which each player arrives through her/his own rational reasoning is mirrored by the one found by any other player. Then, superrationality indicates that the solution to each symmetric game is a profile in the diagonal of the game matrix.

In Hofstadter's words:

> "Any number of ideal rational thinkers faced with the same situation and undergoing similar throes of reasoning agony will necessarily come up with the identical answer eventually, so long as reasoning alone is the

> ultimate justification for their conclusion. Otherwise reasoning would be subjective, not objective as arithmetic is. A conclusion reached by reasoning would be a matter of preference, not of necessity. Now some people may believe this of reasoning, but rational thinkers understand that a valid argument must be universally compelling, otherwise it is simply not a valid argument." (Hofstadter, 1985, p. 746).

Following the lead of Hofstadter, who circumscribed superrationality to symmetric games, we start by considering normal form symmetric games to propose a definition of similarity that extends to non-symmetric games. As a first step we need to find a way to 'symmetrize' non-symmetric games, since in these games it is not clear what could mean 'to make the same choice'. In order to do this we impose a constraint on players' beliefs that amounts to a variant of the well-known symmetry principle in classical bargaining theory. While we do not allow players to perceive others as similar with regards to arbitrary features (such as beauty, trustworthiness, race, etc.) we assume them to reason on the basis of their strategic features. We extend the notion of symmetry to account for situations in which players face the same set of strategies but have different preferences over outcomes.

The plan of the paper is as follows: in the next section we introduce our representation of similarity and provide some examples of strategic interactions in which this notion gives different solutions than Nash Equilibrium. Then we extend this relation to include mixed-strategies. In sections 3 and 4 we formally introduce superrationality and show how it combines with our representation of similarity, yielding interesting insights.

2 A Representation of Similarity

We analyze the notion of similarity in the framework of Game Theory. We start assuming that it can be defined without uncertainty. Recall the definition of games in strategic form:

Definition 1: *Let* $G = \langle I, \{S_i\}_{i\in I}, \{\pi_i\}_{i\in I}\rangle$ *be a game, where* $I = \{1, ..., n)\}$ *is a set of players and a finite set* S_i, $i \in I$ *is the set of actions of player* i. *A profile of actions,* $s = (s^1, ..., s^n)$ *is an element of* $S = \prod_{i\in I} S_i$. *In turn* $\pi_i : \prod_{i\in I} S_i \to \mathbb{R}$ *is player* i*'s payoff function.*

This definition corresponds, in fact, to a *complete information* game. The notion of similarity we introduce here is based only on the objective description of the game. Since it requires to define functions both on the *names* of the players and the *labels* of actions, we will resort to the convention of denoting by s^i_j the action with the j-th label in the set S_i corresponding to player i. Then,

Definition 2: Two players $k, l \in I$ are *similar* if there exist two functions:

1. A permutation $\rho : I \to I$, such that $\rho(k) = l$ and $\rho(l) = k$.

2. A bijection $\phi : \{1, \ldots, |S_i|\} \to \{1, \ldots, |S_{\rho(i)}|\}$ such that for every player $i \in I$ and every $j \in \{1, \ldots, |S_i|\}$ we have:

   (a) $s^i_j = s^{\rho(i)}_{\phi(j)}$,

   (b) $\pi_i(s^i_j, s^{-i}_\kappa) = \pi_{\rho(i)}(s^{\rho(i)}_{\phi(j)}, s^{\rho(-i)}_{\phi(\kappa)})$, where, by a

slight abuse of language we denote $s_\kappa^{-i} =$

$$(s^1_{\kappa_1}, \ldots, s^{i-1}_{\kappa_{i-1}}, s^{i+1}_{\kappa_{i+1}}, \ldots, s^n_{\kappa_n}), s^{\rho(-i)}_{\phi(\kappa)} =$$
$$(s^{\rho(1)}_{\phi(\kappa_1)}, \ldots, s^{\rho(i-1)}_{\phi(\kappa_{i-1})}, s^{\rho(i+1)}_{\phi(\kappa_{i+1})}, \ldots, s^{\rho(n)}_{\phi(\kappa_n)}). \quad (1)$$

To illustrate this notion let us look at the following example of a non-symmetric game with unequal equilibria in the 'diagonal' of S in the case that $S_i = S_j$ for all $i, j \in I$. The reader will easily recognize that the game corresponds to the classic Battle of the Sexes (see Luce and Raiffa, 1957, pp. 90-91). A (reworked) story that usually goes with the game is that players 1 and 2, let us call them Fry and Leela, decide to travel through the Universe separately, and they discuss where they should meet up again. In the example, both Fry and Leela prefer to meet at the end of the day, but have different preferences over where, on what planet, they would like to meet. While Fry prefers Pandora, Leela would rather like to meet in Amazonia, since she hates Pandora. Since they cannot agree on where to meet, they leave the question unsettled. So each has to decide independently whether to fly to Pandora or to Amazonia. The Battle of the Sexes represents their joint dilemma. The game has two pure strategies Nash equilibria in (s^1_1, s^2_1) and (s^1_2, s^2_2).

Example 1

Assume that $S_1 = \{s^1_1, s^1_2\}$ *and* $S_2 = \{s^2_1, s^2_2\}$, *with the following payoff matrix*:

		Player 2	
		s^2_1	s^2_2
Player 1	s^1_1	$(2,1)$	$(0,0)$
	s^1_2	$(0,0)$	$(1,2)$

Table 1: Battle of the Sexes

If we take ρ to be the nontrivial permutation of 1 *and* 2 *(the names of the players) while ϕ is the non-identity over $\{1, 2\}$ (the names of the actions) we have that, for instance,*

$$2 = \pi_1(s_1^1, s_1^2) = \pi_{\rho(1)}(s_{\phi(1)}^{\rho(1)}, s_{\phi(1)}^{\rho(2)}) = \pi_2(s_2^2, s_2^1) \quad (2)$$

This means that, in this game, players 1 *and* 2 *are similar if, when they exchange their names, and under their new aliases choose the actions assigned by the permutation to their original namesakes, the payoffs do not change. Or, to put it in other words, payoffs are invariant with respect to the permutation of the players and strategies.*

Let us notice that this similarity relation exists even if the game is not *symmetric*. According to Maskin and Dasgupta (1986), a game G, in which $S_i = S$ for every $i \in I$, is symmetric if for every permutation $\rho : I \to I$,

$$\pi_i(s_{j_1}^1, \ldots, s_{j_i}^i, \ldots, s_{j_n}^n) = \pi_{\rho(i)}(s_{j_1}^{\rho(1)}, \ldots, s_{j_i}^{\rho(i)}, \ldots, s_{j_n}^{\rho(n)}). \quad (3)$$

That is, a game is symmetric if we can exchange the names of players without modifying the payoffs. Similarity, instead, is a property of players that may have different action sets and thus it requires an extra function from the set of actions of a player to each of those of the players to whom he is similar.[1]

A first question that can be raised is whether the similarity between players has some impact on the solution of a game. The best known solution notion is Nash equilibrium:

Definition 3: *A profile of actions, $s = (s_*^1, \ldots, s_*^n) \in \prod_{i \in I} S_i$ is a (pure strategies) Nash equilibrium if*

[1] The game in Example 1 is not symmetric, even if the action sets are identical.

$\pi_i(s^1_*, \ldots, s^i_*, \ldots, s^n_*) \geq \pi_i(s^1_*, \ldots, s^i_k, \ldots, s^n_*)$, *for every* $i \in I$ *and for every* $s^i_k \in S_i$, $s^i_* \neq s^i_k$.

The following result shows that if every player is similar to all the other players equilibria are exchangeable:

Proposition 1: *Given a game* G *such that* $S_i = S$ *for all* $i \in I$, *if there exist permutations* $\rho : I \to I$ *and* $\phi : \{1, \ldots, |S|\} \to \{1, \ldots, |S|\}$, *such that every player* i *is similar to any other player under them, a (pure strategies) profile* $(s^1_{j_1}, \ldots, s^i_{j_i}, \ldots, s^n_{j_n})$ *is a Nash equilibrium if and only if* $(s^{\rho(1)}_{\phi(j_1)}, \ldots, s^{\rho(i)}_{\phi(j_i)}, \ldots, s^{\rho(n)}_{\phi(j_n)})$ *is also a Nash equilibrium.*

Proof: *Trivial. For any* $i \in I$, $\pi_i(s^1_{j_1}, \ldots, s^i_{j_i}, \ldots, s^n_{j_n}) \geq \pi_i(s^1_{j_1}, \ldots, s^i_a, \ldots, s^n_{j_n})$, *for every* $s^i_a \neq s^i_{j_i}$. *Then, for player* $\rho(i)$ *we have that* $\pi_{\rho(i)}(s^{\rho(1)}_{\phi(j_1)}, \ldots, s^{\rho(i)}_{\phi(j_i)}, \ldots, s^{\rho(n)}_{\phi(j_n)}) \geq \pi_{\rho(i)}(s^{\rho(1)}_{\phi(j_1)}, \ldots, s^{\rho(i)}_b, \ldots, s^{\rho(n)}_{\phi(j_n)})$ *for any* $s^{\rho(i)}_b \neq s^{\rho(i)}_{\phi(j_i)}$.

This result applies immediately to two-player games:

Example 1 **(ctd.)**: The pure strategies Nash equilibria in the off-diagonal Battle of the Sexes are (s^1_1, s^2_1) and (s^1_2, s^2_2). It is clear that under $\rho : \{1, 2\} \to \{1, 2\}$ and $\phi : S \to S$ defined as:

- $\rho(1) = 2$ and $\rho(2) = 1$.
- $\phi(1) = 2$ and $\phi(2) = 1$.

both players are similar to each other. It follows, according to Proposition 1, that each Nash equilibrium obtains applying ρ and ϕ to the other equilibrium.

Let us now consider the following off-diagonal version of the Battle of the Sexes, which would allow us to illustrate a different assignment of the actions ac-

cording to function ϕ.

Example 2: *Assume that, as in Example 1,* $S_1 = \{s_1^1, s_2^1\}$ *and* $S_2 = \{s_1^2, s_2^2\}$*:*
If we take

		Player 2	
		s_1^2	s_2^2
Player 1	s_1^1	$(0,0)$	$(2,1)$
	s_2^1	$(1,2)$	$(0,0)$

Table 2: Off-diagonal Battle of the Sexes

- $\rho(1) = 2$ *and* $\rho(2) = 1$.
- $\phi(1) = 1$ *and* $\phi(2) = 2$.

we have:

$$\pi_1(s_1^1, s_2^2) = \pi_{\rho(1)}(s_{\phi(1)}^{\rho(1)}, s_{\phi(2)}^{\rho(2)}) = \pi_2(s_1^2, s_2^1)$$

and

$$\pi_2(s_1^1, s_2^2) = \pi_{\rho(2)}(s_{\phi(1)}^{\rho(1)}, s_{\phi(2)}^{\rho(2)}) = \pi_1(s_1^2, s_2^1)$$

and thus, Proposition 1 applies, allowing to define one pure strategies Nash equilibrium in terms of the transformation of the other equilibrium through ρ *and* ϕ.

In this game, players 1 *and* 2 *are similar if, when they exchange their names, and under their new aliases choose the same actions as the original namesakes, the payoffs do not change.*

3 Similarity extended to mixed strategies

Let us now introduce an epistemic turn, showing the impact of similarity on the beliefs the agents have

about the others. The representation we choose is the following:

Definition 4: *Given S_i, i's actions set, let $\Delta(S_i)$ be the set of probability distributions over S_i. A mixed strategy of player i is an element $\sigma_i \in \Delta(S_i)$.*

The notion of Nash equilibrium can be extended to mixed strategies:

Definition 5: *A profile of mixed strategies $(\sigma_*^1, ..., \sigma_*^n)$ in a game G is a Nash equilibrium if for every player i, σ_*^i satisfies* [2]

$$E\pi_i(\sigma_*^1, \ldots, \sigma_*^i, \ldots, \sigma_*^n) \geq E\pi_i(\sigma_*^1, \ldots, \sigma_j^i, \ldots, \sigma_*^n)$$

for every $\sigma_k^i \in \Delta(S_i)$, $\sigma_*^i \neq \sigma_j^i$.

Instead of thinking of a mixed strategy of a player i as a randomization over his own actions, the epistemic interpretation is that it summarizes the conjectures or beliefs that the other players have about i. More precisely:

Definition 6:*A belief of player i about the actions to be chosen by the others is given by a probability distribution $p_i(\cdot) \in \Delta(S_{-i})$. The marginal over player j, denoted $p_{j|i}(\cdot)$, is a distribution in $\Delta(S_j)$.*

A well-known result of Aumann and Brandenburger (1995) indicates that a sufficient condition for the existence of a mixed strategies equilibrium is that an *a priori* belief of each player must give positive weight to a state in which the structure of the game and the rationality of the players is mutually known while the

[2] $E\pi_i$ is the expected payoff of i, according to the probabilities defined by the mixed strategies profile.

conjectures about the others are common knowledge. Then, in an equilibrium all the conjectures of the players $j \neq i$ coincide about the choices of i. More precisely, $\sigma^i_*(s^i_k) = p_{i|j}(s^i_k)$ for every j such that $j \neq i$ and for every $s^i_k \in S_i$.

A result equivalent to Proposition 1 ensues for mixed strategies equilibria:

Proposition 2: *Given a game G such that $S_i = S$ for all $i \in I$ in which the players satisfy the conditions of Aumann and Brandenburger, if there exist permutations $\rho : I \to I$ and $\phi : \{1, \ldots, |S|\} \to \{1, \ldots, |S|\}$, such that every player i is similar to any other player under them, a mixed strategies profile $(\sigma^1, \ldots, \sigma^i, \ldots, \sigma^n)$ is a Nash equilibrium if and only if $\sigma^i = \sigma^{\rho(i)}$, where $\sigma^{\rho(i)}$ is a distribution over the set $\{s^{\phi(i)}_{\rho(j)} = s^i_j : s^i_j \in S_i\}$..*

That is, the similarity between all the players extends to their beliefs. When shared by all of them, they can be exchanged without losing their character of *best responses* to each other.

***Example 1* (ctd.)**: *The unique non-degenerate mixed strategies Nash equilibrium in the Battle of the Sexes is $\langle(\frac{2}{3}, \frac{1}{3}), (\frac{1}{3}, \frac{2}{3})\rangle$. Recall that*:

- $\rho(1) = 2$ and $\rho(2) = 1$.

- $\phi(1) = 2$ and $\phi(2) = 1$.

If we take, for instance

$$\begin{aligned}\sigma^1 &= (\text{prob}(s_1^1), \text{prob}(s_1^2)) \\ &= (\frac{2}{3}, \frac{1}{3}), \text{ then } \sigma^{\rho(1)} \\ &= (\text{prob}(s_{\phi(1)}^{\rho(1)}), \text{prob}(s_{\phi(1)}^{\rho(2)})) \\ &= (\text{prob}(s_2^2), \text{prob}(s_2^1)) = \sigma^2.\end{aligned}$$

For another example:

***Example 2* (ctd.)**: *Analogously, the unique non-degenerate mixed strategies Nash equilibria in the off-diagonal Battle of the Sexes is* $\langle(\frac{2}{3}, \frac{1}{3}), (\frac{2}{3}, \frac{1}{3})\rangle$. *Recall that:*

- $\rho(1) = 2$ and $\rho(2) = 1$.
- $\phi(1) = 1$ and $\phi(2) = 2$.

If we take, for instance $\sigma^1 = (\frac{2}{3}, \frac{1}{3}) = \sigma^2$. *It is immediate that* $\sigma^1 = \sigma^{\rho(1)} = \sigma^2$.

4 Superrationality

The Harsanyi doctrine (see Aumann, 1976) prescribes analyzing games on the basis of an universal normative principle of rationality, as to yield a unique (but possibly probabilistic) solution to every fully described decision problem. It states that in a world in which there exists common knowledge of rationality, all players will reason in the same uniquely rational way. But then, such reasoning process may lead to Pareto-inefficient outcomes, as in Harsanyi's analysis of the Prisoner's Dilemma.

Since superrational players start by assuming that there is only one possibly rational outcome, they

strive to reason in a way that leads to it. In this sense it could be thought of as a radical form of the Harsanyi doctrine. Superrational players are like 'puppets of reason', epistemically equipped to think only in terms of joint rationality. Hofstadter did not delve into a philosophical discussion on whether superrational players can do otherwise, for instance thinking counterfactually: 'what if I had done otherwise?' In the case of the Prisoners' Dilemma, such counterfactual leads each player to think of the advantages of defecting. Since superrational thinkers cannot think in such terms, we can argue that they lack agency, since otherwise they would be able, because of their free will, to do otherwise (for a discussion of these topics Kane 1996; cfr. Frankfurt, 1969, for an alternative view which challenges the idea that freedoom requires the possibility of doing otherwise). On the other hand, the notion of superrationality goes beyond the Harsanyi doctrine by adding the condition that a rational solution to any game should be Pareto efficient (Harsanyi and Selten, 1988).

Superrationality is one of several different notions of solution that have been introduced to account for cooperation in the PD. These notions tend to relax the assumption that there exists causal independence between players' actions (for a treatment of this notion see Bicchieri and Green 1997). One of these notions is *magical thinking*, according to which one's beliefs by their own have causal effects in the world (Shafir and Tversky 1992). In the context of interactive games magical thinking implies that my own action somehow influences the action of the other player, perhaps by increasing the likelihood that the other players will choose the same action (Daley and Sadowsky, 2017). Another notion is that of *translucent players*, who can by some means establish whether the other player is

disposed to cooperate or to defect (Gauthier, 1986; Frank, Gilovich, and Regan, 1993; Spiekermann, 2007; Capraro and Halpern, 2019). Evidential Decision Theory provides an alternative way (different from magical thinking) to account for cooperation in the Prisoners' Dilemma (see Ahmed, 2014). Another such concept, which differs both from evidential decision theory and magical thinking, is *perfect transparent equilibrium*, which is based on the idea that the decision of each player depends counterfactually on the decisions of the others (Fourny, 2020).

As said, a common feature of these concepts is the violation of the independence assumption, ruling out the choice of off-diagonal outcomes. Even if the causal dependence among decisions were ruled out, the notion of dependence would reenter the picture in the form of the intrinsic correlation of beliefs or hierarchies of beliefs (Tohmé and Viglizzo, 2019). So, in a way, even if players may choose their actions independently, they might be correlated at the level of the beliefs they hold (about each other). As noted by Brandenburger and Friedenberg (2008), this is an adaptation to game theory of the idea of *common causal principle* of correlation (Reichenbach, 1956).

Superrationality describes a property of profiles in the "diagonal" of a symmetric game. This is a formal requirement to capture the idea that players reason in a "similar" way. From the outcomes in the diagonal, Hofstadter solution selects the unique strictly payoff-dominant one.

Formally this can be expressed in the following definition:

Definition 7: *Let* $G = \langle I, \{S_i\}_{i \in I}, \{\pi_i\}_{i \in I} \rangle$ *be a*

symmetric game. A profile of actions, $s = (s^1, ..., s^n) \in S = \prod_{i \in I} S_i$ *is a superrational solution if:*

1. $s \in diag(S) = \{(s^1, \ldots, s^n) \in S : s^i = s^j,$ *for all* $i, j \in I\}$

2. *For any* $t = (t^1, ..., t^n)$, $\pi_i(s) \geq \pi_i(t)$, *for every* $i \in I$

According to this, the superrational solution is obtained in a symmetric game in which players recognize the situation as one in which they have the same strategies and face the same constraints. The players must therefore conclude that they will reason in the same way, choosing the same action. Having concluded so, they must realize that the only possible outcomes are those in the diagonal. For example, in the Prisoner's Dilemma the superrational solution must be in the set $\{(C, C), (D, D)\}$ of the following matrix:

		Player 2	
		C	D
Player 1	C	(R, R)	(S, T)
	D	(T, S)	(P, P)

Table 3: Prisoner's Dilemma

where $T > R > P > S$.

Given that the first outcome yields a higher payoff, the cooperative solution should be conceived as being the rational choice. As Hofstadter (1985) puts it: "If reasoning guides me to say C, then, as I am no different from anyone else as far as rational thinking is concerned, it will guide everyone to say C" (p. 746). Provided that players have the same choices, payoffs and information (which translates into symmetry of

the game), the cooperative outcome should ensue.

In what follows we analyze how to extend relations of similarity to superrational players.

5 Similarity and Superrationality

As discussed above, superrationality, defined for symmetric games, prescribes that all players must choose the same alternative, constituting thus a profile in the "diagonal" of the payoff matrix. This profile must maximize the payoffs among those in the diagonal. Let us call an outcome that satisfies this condition $\mathcal{SR}_1$.

The existence of $\mathcal{SR}_1$ is ensured by the symmetry of the game. But in Hofstadter's rationale for this solution, the requirement of symmetry among the players is tantamount to their similarity. We will explore here whether the intuition behind superrationality can be extended to non-symmetric games in which all the players are similar.

The extension of superrationality that we will explore here depends on the similarity among agents. We have, first, to define a set of profiles closed under the application of ρ and ϕ. This captures the intuition drawn from superrationality, where the diagonal consists of a class of profiles closed under the relation of symmetry. In this latter case, the second step in the definition of a solution involves the choice of one of those profiles. As indicated at the end of the previous section, the assumption is that, since players are rational (i.e. seek to choose the most preferred option) they must jointly select the dominant profile in the diagonal.[3] We extend this assumption to similar players,

[3]Some critics point out that this ensues from the confusion

who will seek to maximize their payoffs.

In our take we will seek the solution in the *Pareto frontier*, defined as:

$$\Gamma = \{s \in \prod_{i \in I} S_i : \nexists\, s' \in \prod_{i \in I} S_i \text{ such that for all } i \in I,$$
$$\pi_i(s') \geq \pi_i(s), \text{ and for some } j \in I,\ \pi_j(s') > \pi_j(s)\}$$

Then we have:

Proposition 3: If all the players are similar under permutations ρ and ϕ, we have that if

$$s = (s^1_{\kappa_1}, \ldots, s^i_{\kappa_i}, \ldots, s^n_{\kappa_n}) \in \Gamma.$$

then

$$s' = (s^{\rho(1)}_{\phi(\kappa_1)}, \ldots, s^{\rho(i)}_{\phi(\kappa_i)}, \ldots, s^{\rho(n)}_{\phi(\kappa_n)}) \in \Gamma.$$

Examples 1 and 2, of non-symmetric games, illustrate this result. In the latter case, the class of off diagonal profiles belong to Γ, while in the former the elements in the diagonal constitute Γ.

We can introduce a new solution concept, $\mathcal{SR}_2$:

Definition 8: *A $\mathcal{SR}_2$ result is the uniform joint distribution over Γ. That is, each $s \in \Gamma$ has a probability $\frac{1}{|\Gamma|}$.*

---

between the epistemic and the causal abilities of the players (Bonanno 2015). But, as it has been shown in (Tohmé and Viglizzo 2019) if the players are assumed to have what they call a *BK superrational type* (for Brandenburger and Keisler, 2006) they will choose the superrational profile.This assumption builds upon the idea that similar players will share a common type (invariant under ρ and ϕ).

Since Γ selects the profiles in the Pareto frontier, when there are no coincidences, players toss a coin. The $\mathcal{SR}_2$ assures us that they will choose the same mixed-strategy.

This alternative definition can be justified by invoking the *Principle of Insufficient Reason*, according to which no strategy should be assigned a larger probabilistic weight if it is a priori indistinguishable from the others. Sinn (1980) shows that this is a necessary condition for the characterization of players as maximizers of expected payoffs. It is evident that if the players are rational, and if their preferences are indistinguishable in terms of their rationality (e.g., since they are rational they must reason in the same way), then their choices should be indistinguishable. In that case, they should not assign different probabilities to different strategies.[4]

This condition implies that each strategy belonging to Γ is just as likely as any other strategy.

***Example 2* (revisited)**: $\Gamma = \{(s_1^1, s_2^2), (s_2^1, s_1^2)\}$ *and thus the superrational outcome assigns probability* $\frac{1}{2}$ *to each of the profiles in* Γ. *Notice that the expected payoff of each player is* $\frac{3}{2}$, *larger than the worst outcome in the pure Nash equilibria* (1) *and than the expected payoff of non-pure mixed Nash equilibrium* ($\frac{2}{3}$).

An analogous result obtains in Example 1.

Notice that in a symmetric game in which the Pareto optimal profile is in the diagonal, $\mathcal{SR}_2$ coincides with

[4]Notice that the main objection to this Principle, on the basis of the Bertrand Paradox (Shackel 2007), works only in the case of infinite alternatives, but the games considered here have all finite sets of profiles.

$\mathcal{SR}_1$:

Example 3: *Consider the following Hi-Lo game (see Table 1) where* $S_1 = S_2 = S = \{s_1, s_2\}$*:*

		Player 2	
		H	L
Player 1	H	$(2,2)$	$(0,0)$
	L	$(0,0)$	$(1,1)$

Table 4: Hi-Lo Game

By definition, this game is symmetric and has a $\mathcal{SR}_1$ *profile* (H, H)*. On the other hand, players 1 and 2 are similar under* ρ *and* ϕ *such that:*[5]

- $\rho(1) = 2$ *and* $\rho(2) = 1$.
- $\phi(1) = 1$ *and* $\phi(2) = 2$.

It is easy to check that $\Gamma = \{(H, H)\}$ *and so, the* $\mathcal{SR}_2$ *result is the degenerate distribution that assigns probability* 1 *to* (H, H)*, yielding the same outcome as the* $\mathcal{SR}_1$ *profile.*

This suggest the following result:

Proposition 4: *Given a game* G *in which all the players are similar under every pair of permutations* ρ *and* ϕ*, we would have, for every pair of players* $i, l \in I$ *that the expected payoffs under the* $\mathcal{SR}_2$ *solution are such that* $\frac{1}{|\Gamma|}\sum_{s\in\Gamma}\pi_i(s) = \frac{1}{|\Gamma|}\sum_{s\in\Gamma}\pi_l(s)$ *. Furthermore, no individual payoff can be increased without decreasing the payoff of another player.*

[5] ρ is the permutation that yields the symmetry of the game.

Proof: *Since ρ and ϕ make all the players similar, the profiles in Γ support the same payoffs (just in different order), and thus the expected payoff of each individual obtains as the average (with weight $\frac{1}{|\Gamma|}$) of her payoffs in all those profiles. Now suppose that there exists a profile of payoffs that yields a higher payoff for a player i without decreasing the payoff of the other players. This means that this payoff vector does not belong to Γ and thus must obtain combining a pure strategies profile yielding a higher payoff to i. But this contradicts the definition of Γ. Then, $\mathcal{SR}_2$ yields the highest payoff that can be obtained without decreasing that of any other player.*

Note that in Example 3, the concept of Nash equilibrium fails to give us a unique answer. This indetermination clashes with our intuition that the efficient profile (H, H) is the most preferred one. Harsanyi and Selten (1988) incorporate the notion of payoff-dominance to obtain this solution in coordination games with scope for mutual gain. However, it is disputable whether the intuitive answer remains the same if we let risk dominance to enter the picture. In our approach collective rationality as a criterion of rationality is enough to ensure the result. While the intuition favoring this criterion seems strong, there is some evidence that experimental subjects do not play the strategies that would lead to payoff-dominant equilibria if doing so is 'too risky' - the risk being that other players may fail to play their payoff-dominant strategies (Van Huyck, Battalio and Beil, 1990; Crawford, 1991). For instance in the following *Stag Hunt* game, *Hare* "risk-dominates" *Stag* since the former ensures a minimum payoff of 1 instead of 0. Thus, payoff-dominance and the risk dominance solutions do not go hand in hand, and indeed they pull players in different directions:

		Player 2	
		Stag	*Hare*
Player 1	*Stag*	$(2,2)$	$(0,2)$
	Hare	$(2,0)$	$(1,1)$

Table 5: Stag-Hunt Game

It should not be difficult for the reader to check that $\mathcal{SR}_2$ identifies $(Stag, Stag)$ as the solution profile of the game. This discrepancy between risk and payoff dominance, undermines in a certain degree the superrational approach.

6 Conclusion

In this article we presented a formalization of the relation of similarity among players in games, based on the symmetries between them and their strategies. We considered an alternative solution to games like the Prisoner's Dilemma, different from the traditional Pareto-dominated Nash Equilibrium. Since NE does not fare well in such games, that offer *scope for mutual advantage*, it seems theoretically sound to think of alternative notions that could support cooperative outcomes.

The symmetry argument for cooperation (in the Prisoner's Dilemma), has been previously considered in the philosophical literature (Davis, 1977). In this paper we extended it to non-symmetric games. In order to do that, we applied the notion of similarity. We studied its game-theoretical properties, and showed in examples how it works in different strategic interactions. In addition, this paper provided conditions under which the intuition behind superrationality, i.e.,

that the profile of actions picked out by the superrational solution should be payoff-dominant, can be made sense of in games in which all players are similar. We argued, indeed, that Hofstadter's rationale for superrationality can be seen as based on the assumption of similarity among players.

Similarity has also recently started to permeate in the behavioral sciences. Given its conspicuousness in decision making (see Gilboa and Schmeidler, 2001) it provides the grounds for an explanation of how people reduce the complexity of strategic situations (Jehiel, 2005). We believe that providing a clear formal representation of these matters is a task analytical philosophers are apt to undertake, and can help to clarify and rationalize insights in the behavioral sciences.

6 References

AHMED, A. *Evidence, Decision and Causality*, Cambridge: Cambridge University Press, 2014.

AUMANN, R.'Agreeing to disagree', *The Annals of Statistics*, 4 (6), 1236–1239, 1976.

AUMANN, R., BRANDENBURGER, A. 'Epistemic conditions for Nash equilibrium', *Econometrica*, 63(5), 1161–1180, 1995.

BACHARACH, M. 'Interactive team reasoning: A contribution to the theory of co-operation', *Research in Economics*, 53, 117–157, 1999.

BERNHEIM, D. 'A Theory of Conformity', *Journal of Political Economy*, 102(5), 841–877, 1994.

BICCHIERI, C., GREEN, S. 'Symmetry arguments for cooperation in the prisoner's dilemma',

in Holmström-Hintikka, Ghita, Tuomela, R. (Eds.), *Contemporary Action Theory*, Vol. II. (1997). Synthese Library, 229–249.

BONANNO, G. 'Counterfactuals and the Prisoner's Dilemma', in Peterson, M. (Ed.), *The Prisoner's Dilemma* (2015), Cambridge U. Press, 133–155.

BRANDENBURGER, A., FRIEDENBERG, A. 'Intrinsic correlation in games', *Journal of Economic Theory*, 141 (1), 28–67, 2008.

BRANDENBURGER, A., KEISLER, H.J.'An impossibility theorem on beliefs in games', *Studia Logica*, 84(2), 211–240, 2006.

CAMPBELL, R. K. 'The Prisoner's Dilemma and the Symmetry Argument for Cooperation', *Analysis*, 49 (2), 60–65, 1989.

CAMERER, C. F., HO, T-H., CHONG, J-K. 'A Cognitive Hierarchy Model of Games', *Quarterly Journal of Economics*, 119 (3), 861–98, 2004.

CPRARO, V., HALPERN, J. 'Translucent players: Explaining cooperative behavior in social dilemmas', *Rationality and Society*, 31(4), 371–408, 2019.

Chierchia, G, CORICELLI, G. 'The impact of perceived similarity on tacit coordination: propensity for matching and aversion to decoupling choices', *Frontiers in Behavioral Neuroscience*, 9, 202, 2015.

COLMAN, A. M., GOLD, N. 'Team reasoning: Solving the puzzle of coordination', *Psychonomic Bulletin and Review*, 25(5), 1770–1783, 2018.

CRAWFORD, V. 'An "evolutionary" interpretation of Van Huyck, Battalio, and Beil's experimental results on coordination', *Games and Economic behavior*, 3(1), 25–59, 1990.

CRAWFORD, V., HALLER, H. 'Learning how to cooperate: Optimal play in repeated coordination games', *Econometrica*, 58, 571–595, 1990.

DALEY, B, SADOWSKY, P. 'Magical Thinking: A Representation Result', *Theoretical Economics*, 12, 909–956, 2017.

DAVIS, L. H. 'Prisoners, paradox, and rationality', *American Philosophical Quarterly*, 14(4), 319–327, 1977.

DI GUIDA, S., DEVETAG, G. 'Feature-Based Choice and Similarity Perception in Normal-Form Games: An Experimental Study', *Games*, 4(4), 776–794, 2013.

FEHR, E., SCHMIDT, K. 'A Theory of Fairness, Competition, and Cooperation', *The Quarterly Journal of Economics*, 114(3), 817–868, 1999.

FISCHER, I. 'Friend or Foe: Subjective Expected Relative Similarity as a Determinant of Cooperation', *Journal of Experimental Psychology: General*, 138(3), 341–350, 2009.

FLOOD, M, DRESHER, M. 'Some experimental games', Research memorandum RM-789. Santa Monica, CA: Rand, 1952.

FOURNY, G. 'Perfect prediction in normal form: Superrational thinking extended to non-symmetric games'., *Journal of Mathematical Psychology*, 96, 102332, 2020.

FRANK, R. H., GILOVICH, T., REGAN, D. T. 'The evolution of one-shot cooperation: An experiment', *Ethology and Sociobiology* 14, 247–256, 1993.

FRANKFURT, H. 'Alternate Possibilities and Moral Responsibility', *Journal of Philosophy*, 66(23), 829–839, 1969.

GAUTHIER, D. *Morals by Agreement*, Oxford: Oxford University Press, 1986.

GILBOA, I., SSCHMEIDLER, D. *A Theory of Case-Based Decisions*, Cambridge: Cambridge University Press, 2001.

HARSANYI, J. C., SELTEN, R. *A General. Theory of Equilibrium Selection in Games*, Cambridge, MA: The MIT Press, 1988.

HOFSTADTER, D. R. 'Dilemmas for superrational thinkers, leading up to a luring lottery', *Scientic American*, 248(6), 1983.

HOFSTADTER, D. R. *Metamagical themas: Questing for the essence of mind and pattern*, New York: Basic Books, 1985.

JEHIEL, P. 'Analogy-based expectation equilibrium', *Journal of Economic Theory* 123, 81–104, 2005.

KANE, R. *The significance of free will*, Oxford: Oxford University Press, 1996.

LUCE, D., RAIFFA, H. *Games and decisions: introduction and critical survey*, New York: Wiley, 1957.

MUSSWEILER, T., OCKENFELS, A. 'Similarity increases altruistic punishment in humans', *Proceedings of the National Academy of Sciences*, 110(48), 19318–19323, 2013.

MYERSON, R. B. *Game Theory: Analysis of Conflict*, Boston, MA: Harvard University Press, 1991.

NASH, J. F. 'The Bargaining Problem', *Econometrica*, 18(2), 155–62, 1950.

RAPOPORT, A. *Fights, games, and debates*, Ann Arbor: University of Michigan, 1960.

REICHENBACH, H. *The Direction of time*, Berkeley: University of Los Angeles Press, 1956.

ROSS, D. *Philosophy of Economics*, New York: Palgrave Macmillan, 2014.

RUBINSTEIN, A., SALANT, Y. '"Isn't everyone like me?": On the presence of self-similarity in strategic interactions', *Judgment and Decision Making*, 11 (2), 168–173, 2016.

SHACKEL, 'Bertrand's Paradox and the Principle of Indifference'. *Philosophy of Science*, 74 (2), 150–175, 2007.

SHAFIR, E., TVERSKY, A. 'Thinking Through Uncertainty: Nonconsequential Reasoning and Choice', *Cognitive Psychology*, 24(4), 449–474, 1992.

SINN, H. W. 'A rehabilitation of the principle of insufficient reason', *Quarterly Journal of Economics*, 94(3), 493–504, 1980.

SPIEKERMANN, K. 'Translucency, assortation, and information pooling: how groups solve social

dilemmas', *Politics, Philosophy and Economics*, 6, 285–306, 2007.

TOHMÉ, F., CATERINA, G., GANGLE, R. 'Local and global optima in decision-making: a sheaf-theoretical analysis of the difference between classical and behavioral approache', *International Journal of General Systems*, 46(8), 879–897, 2017.

TOHMÉ, F., VIGLIZZO, I.'Superrational types', *Logic Journal of the IGPL*, 27(6), 847–864, 2019.

VAN HUYCK, J. B., BATTALIO, R., BEIL, R. 'Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure', *American Economic Review*, 80(1), 234–48, 1990.