Hamilton Meets Causal Decision Theory

Johannes Martens

Hamilton Meets Causal Decision Theory

Published in *Studies in History and Philosophy of Biol & Biomed Sci* 77 (October 2019)

Johannes Martens

CNRS/SND

Abstract

In this paper, I contrast two mathematically equivalent ways of modeling the evolution of altruism, namely the classical inclusive fitness approach and a more recent, "direct fitness" approach. Though both are usually considered by evolutionists as mere different ways of representing the *same* causal process (i.e. that of kin selection), I argue that this consensus is actually misleading, for there is a fundamental *ambiguity* concerning the causal interpretation of the DF approach. Drawing on an analogy between the structure of inclusive fitness theory and that of causal decision theory (Stalnaker 1972), I show that only the inclusive fitness framework can provide us with a proper and unambiguous causal partition of the relevant variables involved in the evolution of altruism.

I] Introduction

Explaining the evolution of altruistic behaviours has been and remains one of the major challenges of evolutionary biology. In general, "altruism" refers roughly to any trait that both incurs a cost to the bearer and provides some benefits to one or several individuals, where the cost and benefits are defined in terms of the expected reproductive success of the individuals. Biologists and philosophers have long disagreed about the exact meaning of this definition, especially in what concerns the characterization of the cost of altruism, and these debates are still running nowadays (e.g. Kerr *et al*. 2004, Lehmann et Keller 2006). However, I will focus exclusively in this paper on those altruistic behaviours that involve an absolute fitness cost for the individuals. These traits are typically described as strongly altruistic (Wilson 1979)—though, for the sake of simplicity, I will use the terms "altruism" and "strong altruism" interchangeably.

What distinguishes altruism from other kinds of social behaviour is that it seems to involve a true fitness sacrifice from the viewpoint of the altruist. Understood in this (strong) sense, altruism thus raises a familiar problem for the theory of natural selection, namely: how can a trait that diminishes the fitness of its bearers have evolved so as to give rise to much of the sophisticated behaviours and organizations that we commonly observed in the biological world? Nowadays, this question is no longer a puzzle for the evolutionists, for it is widely admitted that strong altruism can (at least in principle) evolve provided that there are positive correlations in the populations, that is, provided altruists tend to interact more than average with other individuals sharing their predisposition for altruism. However, a persisting problem is to determine the *best* way of modeling these correlations, especially when one is trying

to explain their role in the etiology of altruistic traits; and this is precisely the point I will address in this paper.

There are currently two ways of modeling the fitness structure of (strongly) altruistic traits in the field of social evolution theory, namely the inclusive fitness theory, based on the analogy of organisms-as-maximizing-agents (Grafen 2006, West and Gardner 2013, Birch 2016, Martens 2017), and a recent alternative, known as the "direct fitness" approach. Both approaches rely on the representation of a biological population in terms of "actors" and "recipients", and are considered as mathematically equivalent (Taylor *et al.* 2007; West *et al.* 2007), at least in all the relevant cases where strong altruism can evolve. Multilevel selection theory has often been opposed to inclusive theory as a distinct alternative, but their formal equivalence (Hamilton 1975; Leigh 2010; Marshall 2011) is now well-admitted. So, in this paper, I will implicitly consider (without further arguments) that multilevel selection is just another possible name for the process of kin selection described by the inclusive fitness approach (following Nunney 1985), and allegedly by the direct fitness approach.

The inclusive fitness approach (here abridged the IF approach) explains the evolution of altruism in terms of the indirect benefits that a focal altruist—the actor—provides to her social partners (or recipients), weighted by the proper relatedness coefficients (Hamilton 1964, 1970). In this approach, altruism is defined as a costly behaviour in absolute fitness terms, which can only evolve when some sufficient genetic similarity between the actor and the recipient(s) ensures that the beneficiaries will also be the carriers of the altruistic gene. In contrast, the direct fitness approach (here abridged the DF approach) proceeds by calculating the effect on the reproductive success of a focal altruist of the behaviour of all her social partners, weighted by a correlation coefficient—which, of course, must be high enough in order to altruism to

evolve (Taylor and Frank 1996). Thus, according to this perspective, altruism evolves when the cost incurred by the focal altruist happens to be compensated by a sufficient quantity of correlated direct benefits she receives from her altruistic partners.

Currently, most evolutionists consider the DF approach as a mere alternative way of modeling the same causal process than the one represented by the IF approach, namely kin selection (Taylor *et al.* 2007; West *et al.* 2007). The only noticeable difference, they argue, is strictly methodological, and comes from the fact that the DF approach provides a more convenient way of deriving the specific conditions for the evolution of altruism in a large number of situations (as, for instance, when the members of the population belong to different classes with different reproductive values, such as male or female). Thus, regardless of these methodological subtleties, no loss of information should be observed when one is shifting from the IF to the DF formulation, and vice versa.

As I will show in this paper, however, this common view is actually misleading, for the predictive equivalence of the IF and the DF approaches hides an important causal discrepancy. The latter can be summarized as follows. According to the IF approach, altruism can only evolve through the indirect benefits provided *to* the partners, for the altruists incur an absolute cost which is not compensated by any direct benefits; yet, the DF approach is not as much specific, and only implies that altruism will evolve when the correlated benefits (expected from the partners) are sufficiently high to overcome the cost of altruism. Hence, unlike the IF approach, the DF approach does not specify whether the altruists suffer a *net* direct fitness cost due to their behaviour, and leaves open instead the possibility that altruism evolves through the correlated benefits provided by the partners, rather than through the indirect benefits (provided to the partners); and it is precisely this ambiguity that I intend to remove in this paper.

Part of the reason why this causal difference matters is that there is some enduring confusion concerning the proper characterization of altruism in evolutionary biology, especially concerning the distinction between altruism and reciprocal behaviour (or reciprocity). Reciprocal behaviours evolve primarily due to the direct expected benefits they provide to the cooperative individuals, in contrast with altruistic behaviours, which involves a sacrifice from altruistic individuals—as observed in the worker caste of eusocial insects. However, in many situations, altruism and reciprocal behaviour look identical prima facie, and are easily confused; that is why it is essential to settle on a general criterion that allows us to distinguish between these two types of social behaviour.

In order to determine which of the correlated benefits expected *from* the partners (DF approach) or of the indirect benefits provided *to* the partners (IF approach) cause the evolution of altruism, I will rely in this paper on a formal analogy between the logical structure of the IF perspective and that of causal decision theory (Gibbard and Harper 1981). In rational choice theory, the expected utilities of social strategies are typically modeled using conditional probabilities, which measure the amount of correlations between the individuals within games (Skyrms 1994, 1996). Thus, when one shifts from rational choice theory to evolutionary game theory, the DF approach appears as a simple statistical counterpart of such a probabilistic representation. Yet, against this representation, I will suggest an alternative model for the expected fitness of social strategies, directly inspired by a model of expected utility first proposed by Stalnaker (1972), which aims at representing the proper causal structure of behaviours. Using this alternative framework, I will then show that the IF approach, by focusing on the sole phenotypic impact of a focal altruist's behaviour on the fitness of her

partners, accounts automatically for the causal efficacy of the altruistic strategy (contrary to the DF approach).

The general structure of this paper is the following. In section II, I describe in broad outline the formal relation between the IF and the DF approaches. I also argue that, despite their predictive equivalence, both of these approaches may lead to incompatible descriptions of the process governing the evolution of altruism. Second, I discuss (section III) the main possible ways of modeling the evolution of altruism. To this end, I rely on Skyrms' treatment of the parallel (1994) between the representation of correlations in rational choice theory and in evolutionary game theory, by focusing on the well-known situation of the two-player prisoners' dilemma game. In section IV, Newcomb's problem is introduced through the discussion of both the correlated prisoner's dilemma in rational choice theory and its resolution by causal decision theory. In section V, I argue that strong structural affinities exist between the classical version of causal decision theory (Stalnaker 1972; Lewis 1979, 1981) and the IF approach. In particular, I show that substituting probabilities of counterfactuals to conditional probabilities in the classical model of expected fitness (which corresponds to the DF approach) allows us to take into account the causal efficiency of evolutionary strategies, and removes any ambiguity concerning the causal role of the correlated benefits expected from the partners in the evolution of altruism. Lastly, I consider (section VI) an important objection against the IF view, namely the claim that altruists, in many settings, control assortment—and therefore control indirectly the expected direct benefits they receive from their partners (Rosas 2010).

II] Inclusive fitness vs. direct fitness approaches

IF theory has long been, by far, the most popular approach of the evolution of altruistic characters among biologists (Lehmann and Keller 2006; West *et al*. 2007). As is well-known, it has its root in a couple of papers by Hamilton in 1964, who was the first to demonstrate that, despite the sacrifice involved in carrying out altruistic behaviours, such traits can nevertheless evolve if the quantity of benefits produced is allocated to individuals with a sufficiently high probability—compared to a random individual of the population—of sharing with the altruist the same genetical disposition. His argument is now famous under the form "rb – c > 0", that expresses in precise terms a general condition for the evolution of altruism. In this expression, c refers to the absolute expected cost incurred by the altruist, b to the expected[1] benefits (i.e. the "extra" number of offspring) received by her relatives as a consequence of her behaviour, and r to the coefficient of relatedness measuring the degree of genetic similarity between the altruist and the recipients at the relevant loci, i.e. at the loci determining the expression of the altruistic phenotype (Hamilton 1970, 1975). The quantity rb – c resulting from a slight change in the altruistic phenotype of an individual corresponds to what Hamilton called the "IF effect" of altruism, and corresponds to a quantity that is maximized by organisms in social contexts, assuming additive payoffs (Grafen 2006). The process by which altruism evolves under the action of natural selection, when rb – c > 0 and rb > 0 > (– c), is called kin selection (Maynard Smith 1964).

[1] The coefficients b and c are both statistical quantities which correspond, respectively, to the regression of the fitness of the partner on the phenotype of the focal individual, and to the regression of the fitness of the focal individual on her own phenotype. But throughout this paper, I will suppose that the populations to which the rule can be usefully applied are sufficiently large, so that the average reproductive gain and loss denoted by b and c could be envisaged as a good approximation of their expected value.

Hamilton claimed that IF was a good way for explaining the evolutionary success of social characters, especially altruism, as it provides a very natural partition of the two different pathways, direct and indirect, involved in the transmission of these traits. However, there exists another possible way of accounting for the cost and the benefits of altruism, which consists in computing the net expected number of offspring that an individual should gain in behaving a little more altruistically—a quantity that Hamilton (1964) called the "neighbor-modulated fitness". Due, in part, to the overwhelming success of the IF theory, this approach has long been neglected by most of the evolutionists interested in the evolution of social behaviour. But thanks to its formal elaboration in a seminal paper by Taylor and Frank (1996), it has progressively emerged as the "preferred approach of theoreticians" (Taylor *et al.* 2007, 301), though mainly for reasons of mathematical convenience. Taylor and Frank dubbed their method the "direct fitness (DF) approach", in that it accounts for the expected effects of the social environment on the reproductive success of a random focal altruist (the "neighbor-modulated" part of the total reproductive success), instead of indirect benefits. I will follow their terminology.

Putting aside the technical details, the nature of the relation between these two accounting systems can be quite easily grasped. Actually, the main difference between the IF and the DF approaches lies in the way they account for the patterns of fitness interaction that determine the direction of natural selection on altruistic characters.

On the one hand, the IF approach can be characterized as an *actor-centered view*, which keeps track of all the effects b_1, b_2, …, b_k resulting from the action of a single focal altruist—the actor—on the fitness of the k recipients, weighted by the proper relatedness coefficients r_1, r_2, …, r_k, plus the effect on her direct fitness, (– c).

According to this perspective, altruism will thus evolve if Hamilton's rule is satisfied (formula 1), where the LHS of the inequality is the IF of the focal actor:

$$(1)\ \sum_{k} r_k b_k - c > 0$$

In contrast, the DF approach is primarily a *recipient-centered view*, which proceeds by calculating the expected fitness effect on a focal recipient of the behaviour of all her k social partners (i.e. weighted by the proper correlation coefficients a_1, a_2, …, a_k), plus the direct effect on her fitness:

$$(2)\ \sum_{k} a_k b_k - c > 0$$

Here, the LHS of the inequality is the net DF (or neighbor-modulated fitness) of the focal recipient.

In this context, any interaction is basically defined as a singular causal relation between two basic units, namely: an actor, whose phenotype affects one or several components of fitness, and a recipient, whose fitness can be affected by several phenotypes. These two categories, however, are not mutually exclusive (e.g. a focal individual affecting her own reproductive success through her phenotype, as measured by (– c), will both count as an actor and a recipient[2]), but correspond simply to distinct functional roles that help to specify the causal patterns existing in a given population, relevant to a particular social trait. Thus, IF and DF are simply two different ways of accounting for these causal relations.

[2] This point is merely a trivial consequence of the fact that a focal individual has a relatedness of 1 to himself.

As one can see from the comparison of (1) and (2), the general form of the rule for the evolution of altruism one obtains from the DF approach is the same as that of the original Hamilton's rule. But this formal similarity is somewhat misleading, because in the two cases it has to be read differently. Indeed, while $r_k b_k$ measures the quantity of indirect benefits to the altruist in the IF framework (where r_k is the regression coefficient of the genotypes of the k recipients on the phenotype of the focal individual i), each of the a_k coefficients stands for a measure of the regression of the phenotypes of the k social partners on the genotype of the focal individual i. So according to the DF perspective, the a_k coefficients merely reflect the extent to which the more altruistic individuals tend to have more altruistic partners, *regardless* of the latter's genotypes. This structural distinction is summarized in the figure 1, where arrows represent causal influences from actor's phenotype(s) to recipient's fitness(es), and thus recipient's genotype(s).

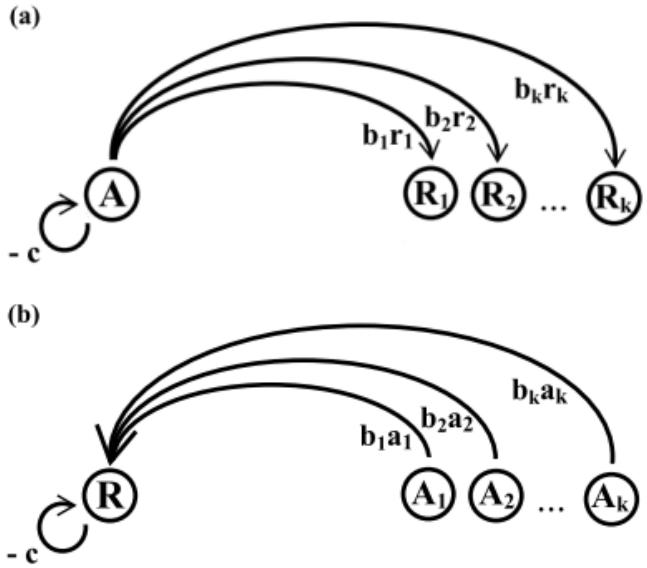

Figure 1. Actor vs. recipient centered view: two ways of accounting the fitness effects of altruism (A stands for "actor" and R stands for "recipient").

Despite this discrepancy, however, most of the evolutionists using the DF fitness approach seem to consider that there are no true conceptual differences between the two methodologies (West *et al.* 2007; Taylor *et al.* 2007), which they tend to see as two *equally valuable* ways of describing the same process, namely that of kin selection. Taylor *et al*. (2007), in particular, clearly stress that the difference between formulae (1) and (2) are "purely technical" and "derives from the re-indexing" (p.304) of the expressions $a_k b_k$ and $r_k b_k$ in the LHS of these inequalities. Of course, one can always find some situations where the a_k and the r_k will not be equivalent. But for these authors, "this arises from (interesting) attempts to extend the methods beyond their normal range of applicability, e.g. to fitness interactions between species" (Taylor *et al*. 2007, p.302). And indeed, this argument makes sense, to the extent that the DF method was first designed by Taylor and Frank (1996) to supply IF theory with nothing more than a better way of coping with the great diversity of kin selection scenarios. So, in a word, the common view is that the general form of the rule (2) obtained through the DF approach may be interpreted in IF terms without loss of any relevant information.

However, this widespread consensus regarding the nature of the conceptual relationship between the IF and the DF approaches is ultimately misleading, for the DF approach is actually *far more general* than the IF approach. The main reason for this greater generality comes from the fact that the DF approach says nothing about the cause of the correlated benefits returned to the focal recipient. It could, for instance, be a common cause, as when the recipient is kept close to her nearest relatives through some external mechanism, or through some mechanism of phenotypic recognition that binds the partners together (or any combination of the two); but it could also be a more

proximal cause, as when the behaviour of the focal recipient is itself the cause (though more or less distant) that, ultimately, determine the behaviour of her partners.

In this latter case, the correlated benefits expected from the partners, i.e. $\sum_k a_k b_k$, would then be the most important cause of the evolution of the social behaviour—and the latter would thus be a form of reciprocal behaviour (Trivers 1971). But here lies precisely the heart of the problem, namely: even if the DF approach doesn't imply logically that the correlated direct benefits are the primary cause of the evolution of altruism (as everyone knows, correlation doesn't imply causation), it still fails to *disambiguate* whether these benefits could be construed legitimately (or not) as the ultimate cause in the evolution of altruism.[3] Thus, while statistically correct, the DF approach is actually quite confusing.

Currently, the causal interpretation of the correlated benefits term (expected from the partners) is not held by the whole majority of the biologists using the DF methodology—who merely emphasize the formal aspect of the equivalence between the IF and the DF frameworks, without paying any real attention to the problem raised by their respective causal interpretations. But it has its advocates, both among biologists and among philosophers. Thus, Fletcher and Doebeli (2009, 17), for instance, reject "the erroneous claim that altruism can only evolve via indirect fitness", and defend a generalized DF approach for explaining the evolution of altruistic traits. Rosas (2010) is also quite explicit in his rejection of the IF framework, though not so much on the causal role of correlated benefits in the evolution of altruism:

> "If you think that the altruistic gene will spread in spite of the fact that donors are losers, you are victim of an *illusion*: although some donors will be losers, they are not losers on average, and cannot be, if altruism evolves. […] The

[3] As an example, just consider the fact that the same partition (i.e. the LHS of equation 2) can be used in order to model both the evolution of (strong) altruism and that of mere reciprocity.

> invitation, therefore, is to see the personal or direct fitness approach suggested by the DF methodology as the fundamental and more general perspective." (Rosas 2010, p.6, my emphasis)

Of course, such a causal interpretation of the DF approach, if it turned out to be the right one, would have radical implications for our understanding of kin selection. In particular, it would make obsolete the whole Hamiltonian picture relying on the central role of indirect benefits as the primary cause of the evolution of altruistic characters, as well as the "orthodox" conception of altruism as a kind of "self-sacrificing behaviour". And strong altruism, in this perspective, would not fundamentally different from any other kind of biological adaptation, as its evolution could be simply explained by the fact that it maximizes the expected reproductive success of the individual altruists.

There is, however, another interesting aspect of the problem that remains to be envisaged. Given, indeed, that it is now well-admitted that the DF methodology, as West *et al*. (2007) notice, "has revolutionized social evolution theory, providing a simpler method that produces more general models, compared with the inclusive fitness approach" (425), then why don't these theorists merely abandon the old IF perspective in favour of a generalized DF approach? Taylor *et al*. (2007), in the conclusion of their paper, provide an interesting element of answer. According to them:

> "The popularity of direct fitness in the theoretical literature in recent years reflects the fact that it is often mathematically more natural to formulate. But the inclusive fitness paradigm continues to have a powerful presence, no doubt because it positions the modeller as an agent choosing behaviour to maximize fitness. This 'individual as maximizing agent' analogy (Grafen 1999) allows us

> to put ourselves in the position of an individual organism and ask: how can I maximize my inclusive fitness? Not only does this constitute a powerful theoretical construct, but it is also a natural question for us humans to ask, as in our day-to-day lives our behavioural decisions are typically optimal, albeit with regard to complex payoff functions." (Taylor *et al*. 2007, 308)

Thus, according to these authors, the current pervasiveness of the IF approach—despite the greater mathematical convenience of the DF approach—would be mainly explained by its *heuristic* advantage, that is, by its structural similarity with our daily way of making economic decisions (Grafen 2006; West and Gardner 2013).

Taylor *et al*. (2007), in my view, are fully right to insist on the importance of the relation between the IF perspective and the analogical stance of a maximizing agent in what concerns its heuristic power. But their previous explanation is actually insufficient; for as I will show in the next sections, a stronger argument can be made in favour of the IF interpretation. More specifically, I will show that the IF approach—due to its structural link with the perspective of a focal actor thought as a maximizing agent—provides us not only with a better heuristic than the DF approach, but also with a proper and unambiguous *causal decomposition* of the fitness constituents involved in the evolution of altruism.

III] From utility to fitness: modeling the evolutionary success of altruism

There is actually a common fitness structure beyond the great diversity of strongly altruistic traits, which corresponds to the general form of the one-shot prisoner's dilemma game. In order to represent this structure through a unique payoff matrix, I will suppose a general biological setting where the individuals interact in pairs, where

the costs and benefits are additive. Clearly, neither of these two assumptions is required for the application of either the DF or the IF approaches (Gardner *et al.* 2011); but when the assumption of additivity is dropped, many troubles arise for the causal interpretation of the results (Okasha and Martens 2016; Okasha 2016). The corresponding fitness structure is given by figure 2.

	Altruism (A)	Selfishness (S)
Altruism (A)	b – c	– c
Selfishness (S)	b	0

Figure 2. The one-shot prisoner's dilemma.

Given this general setting, what could be the more appropriate way of representing the respective fitnesses of both altruist and selfish individuals? As it is well-known, the success of a phenotypic strategy in standard evolutionary game theory (Maynard Smith 1982) is most often formalized according to the classical model of expected utility that is found in classical rational choice theory (e.g. Savage 1954), and which consists in computing its expected reproductive value across all relevant kinds of social environments. In classical game theory, for instance, the utility of an act is characterized as a probability-weighted average of the utilities associated with each possible outcome, given that the outcomes also depend on the strategy X_i chosen by the partner (von Neumann & Morgenstern 1944).[4] One can therefore represent the expected utility of an act through the general expression:

[4] In this paper, I will use the terms "act" and "strategy" interchangeably.

$$(3)\ EU(Act) = \sum_{i} P(X_i).U(Act(X_i))$$

Transposed to our simple evolutionary model, these formulae generate the following expressions for the expected fitness of altruism EW(A) and selfishness EW(S):[5]

$$(4)\ EW(A) = P(A)(b - c) + P(S)(- c)$$

$$(5)\ EW(S) = P(A)(b) + P(S)(0)$$

Here, P(A) and P(S) are respectively the (unconditional) probabilities that the focal individual interacts with a partner of type A (altruist) and with a partner of type S (selfish), which are equal to the respective frequency of A and S individuals in the global population, provided that the population is sufficiently large. The fitness payoffs are the same as those given in figure 2.

A well-known implication of the analogy between utility and fitness is that it allows the modeler to think about the outcome of natural selection from the viewpoint of a focal member of the population *as if* the latter was a rational agent. The key idea of this view is that what would be chosen by a rational agent seeking to maximize her expected fitness should also correspond to the strategy favoured by natural selection in the whole population, ceteris paribus. This method, dubbed by Sober (1998) the "heuristic of personification", and also known as the maximizing agent analogy (Grafen 1999, 2006), is probably one of the most pervasive heuristics used in the field

[5] The expected fitness of altruism refers here to the expected fitness of an *individual* altruist. Thus, one may think of altruism as a variable taking a value of either 1 or zero depending on whether the individual is an altruist or a non-altruist.

of social evolution theory. Of course, the organisms modeled in this way are not at all rational agents, as they cannot even "choose" between their phenotypes. Thus, it is crucial to keep in mind that the purpose of this analogy is merely to determine the outcome of natural selection in biological populations, where natural selection acts so as to maximize the representation of genotypes that appear to be correlated (in average) with higher fitness values for individuals.

Despite its heuristic advantages, this method is also known to give rise to wrong predictions when it is applied in an evolutionary context where there are correlations in biological populations (Skyrms 1994, 1996; Sober 1998). Skyrms, actually, was the first to clearly stress the limitations of this analogy, in his paper entitled "Darwin meets the Logic of Decision" (1994), in which he showed that the use of such an analogical reasoning cannot correctly predict the outcome of natural selection in social context when the assumption of random interaction between individuals is dropped. Indeed, a rational agent seeking to maximize her own expected utility would always choose to behave selfishly in a standard one-shot prisoner's dilemma, because of $EU(A) - EU(S) > 0$; and, accordingly, the plain transposition of this reasoning in an evolutionary context should systematically lead us to predict the evolution of selfishness; yet, as Skyrms rightly noticed, this is an incorrect result, for biological altruism *can* evolve provided that positive correlations are sufficiently high in the populations.

Interestingly, Skyrms argued that, in presence of correlations, the expected fitnesses of each strategy could nevertheless be computed in evolutionary game theory according to one class of models in rational choice theory, which use conditional probabilities as weights of the utility values rather than simple, unconditional probabilities. Skyrms, in this paper (1994), focused more especially on the model of

conditional expected utility developed by Richard Jeffrey in his book *The Logic of Decision* (1983), for Jeffrey allowed explicitly for the possibility that the act chosen by an agent might influence the probability of the states which determine her payoffs.

In practice, Jeffrey's model consists in using as weights in the calculation of the expected utility of an act the probabilities of each state X_i *conditional on* the action in question, namely $P(X_i|Act)$, and not the unconditional probabilities of each state, $P(X_i)$. Most often, however, the most interesting cases are those in which the states are simply the acts of the other player, and in those cases, $P(X_i|Act)$ has to be read as a measure of conditional information about the strategy X_i chosen by the partner, given the strategy A chosen by the agent. In this perspective, the expression for the expected utility of an act becomes:

$$(6)\ EU(Act) = \sum_i P(X_i|Act).U(Act\&X_i)$$

Applied to our simple evolutionary setting exposed at the beginning of this section, this model seems now to give the correct measures for the expected fitnesses of altruism and selfishness, allowing for any amount of positive (or negative) correlations in the population, namely:

$$(7)\ EW(A) = P(A|A)(b - c) + P(S|A)(- c)$$

$$(8)\ EW(S) = P(A|S)(b) + P(S|S)(0)$$

In these formulae, each $P(X_i|Y)$ measures the probability that an individual of type Y interact with an individual of type X_i (i.e. either A or S). So, subtracting (8) from (7),

we can easily recover a direct and general condition for the evolution of altruism in the biological context of one-shot prisoners' dilemmas within correlated pairs:

$$(9)\ [P(A|A) - P(A|S)]b - c > 0$$

This condition, unlike the one that can be derived from (4) and (5), gives us the right predictions for determining the outcome of natural selection in presence of correlations. But most importantly, it also expresses a general DF condition for the evolution of altruism. For here, P(A|A) – P(A|S) embodies a general measure of *phenotypic* correlation within pairs, and one that precisely reflects the DF change between the altruist condition and the selfish condition, independently of any pattern of relatedness.[6] Skyrms, in his works, suggested an interesting way of thinking about the meaning of this general formula. In particular, he noticed that, in the special case of perfect autocorrelations (i.e. where P(A|A) = P(S|S) = 1), natural selection enforces a sort of *Darwinian version of Kant's categorical imperative*, which he stated as: "Act only so that if others act likewise fitness is maximized" (Skyrms 1994, p.525). Most of the time, however, correlations are not perfect in natural populations, and the categorical imperative must then be weakened, allowing for any value of the conditional probabilities, as well as for any group size (Bergstrom 2002, for this reason, speaks of this condition as analogous to a "semi-Kantian imperative"). But, once taken in its weak (hence more general) version, this imperative provides a general and predictively accurate version of the heuristic of personification, which can be applied to predict the outcome of social selection in presence of correlations.

[6] Formula (9), indeed, is merely a special instance of condition (2) where k = 1 and where the strategy set is discrete (A or S) rather than continuous (i.e. admitting different degrees of altruism).

IV] Correlated prisoner's dilemma as a Newcomb's problem

There is an important difficulty that is faced by Jeffrey's model when it is interpreted as a system for thinking about rational decision, namely: it leads to predictions that are at odd with the standard dominance principle each time that it is applied to Newcomb's problems. As I will show in the next section, this aspect of Jeffrey's model also turns out to have some decisive repercussions on its use for modeling the expected fitnesses of social strategies within evolutionary theory, and especially the expected fitness of altruism.

Roughly, a game-theoretic situation can be considered as a Newcomb problem when the acts that count as inferior according to the dominance principle are correlated with some efficient outcomes that they do not yet causally promote. The most explicit instance of Newcomb problem is the case of the prisoner's dilemma with a near psychological twin, first described by Nozick (1969), and then popularized by Lewis (1979). As is well-known, altruism is always a strictly dominated strategy in the prisoner's dilemma, for whatever the strategy chosen by its partner (i.e. altruism or selfishness), each player does better if acting selfishly. Thus, following the dominance principle, selfishness it the only rational choice, though both of the players would have eventually fared better in acting altruistically rather than in acting selfishly. In the prisoner's dilemma with a near psychological clone, each player knows that choosing to act altruistically constitutes a highly reliable evidence that the other will act in the same way, an evidence which is precisely captured by the close-to-one values of P(A|A) and P(S|S). Thus, in this context, altruism becomes the strategy that maximizes Jeffrey's expected utility. Yet, one cannot conclude from this evidence that altruism is the truly rational option here. For players are supposed to be prisoners, making their

choice in two separate cells, and without any possibility of communicating. This implies, crucially, that none of their acts can influence the decision of the other.

In this case, the conditional probabilities of Jeffrey's model reflect merely "evidential relevance with causal independence" (Skyrms 1994, p.506) which is precisely the hallmark of spurious correlations (reflected by the very high values P(A|A) and P(S|S)). But clearly, a decision relying exclusively on spurious evidences and without any regard for the causal efficacy of the allowed strategy will hardly be deemed a rational one. So, for this reason, the maximization of Jeffrey's expected utility cannot be considered as a satisfying principle for rational choice.

Jeffrey-like models are not, however, the only way of computing the expected utility of an act. In particular, there is, in rational choice theory, a famous way of reconciling the dominance principle with the principle of expected-utility maximization, allowing to get rid with spurious evidence. This solution consists in computing the expected utility of an act using probabilities of counterfactual $P(Y \rightarrow X_i)$, rather than conditional probabilities $P(X_i|Y)$—such as in Jeffrey's model. This solution, first suggested by Stalnaker (1972) in a letter to Lewis, has been popularized under the name of *causal decision theory* (Gibbard and Harper 1981; Joyce 1999), though not all current versions of this theory resort to the use of counterfactuals. In Stalnaker's version of this theory, the corresponding formula for the expected utility of an act is:

$$(10)\ EU(Act) = \sum_i P(Act \rightarrow X_i).U(Act\&X_i)$$

Contrary to conditional probabilities, the main specificity of the probabilities of counterfactuals relies in their exclusive sensitiveness to the causal influence between

events. So, given two events X and Y, the following equalities hold each time that the correlations between X and Y are spurious ones:

(11) $P(Y \rightarrow X) = P(\sim Y \rightarrow X) = P(X)$

Accordingly, in Stalnaker's version of causal decision theory, altruism is never the rational option in the one-shot prisoner's dilemma, for behaving altruistically does not have any causal impact on the choice made by the partner, so that:

(12) $P(A \rightarrow A) = P(S \rightarrow A) = P(A)$

All of the versions of causal decision theory are not, of course, committed to the use of probabilities of counterfactuals (e.g. Cartwright 1979; Skyrms 1980; Lewis 1981). However, all have in common the commitment to the idea that only *causally relevant* evidences should be taken into account in the calculation of expected utilities. Using probabilities of counterfactuals is just a convenient way for achieving this purpose.

Coming back to the application of Jeffrey's model in evolutionary dynamics (the so-called "Darwinian categorical imperative"), we are now left with two options. A first possibility would be to accept the objection to Jeffrey's model as a model of rational choice, while denying its importance when the latter is used in the context of evolutionary thinking. This is the position held by Skyrms himself, who argues that "results problematic for Jeffrey's theory as a theory of rational choice make perfect sense in the context of population dynamics" (Skyrms 1994, p.504). Skyrms, more precisely, accepts the core idea of causal decision theory, but claims that in presence of correlations, natural selection and rational deliberation can still "part ways", for

natural selection in this context may lead, as we have seen, to the evolution of altruism, which is a strictly dominated strategy.

However, Skyrms is not really explicit about the meaning of Jeffrey's model when used in an evolutionary context. A reasonable way of explaining its role would be to say that, while this model does not constitute a proper *description* of the process of rational deliberation, it nevertheless provides us with a correct *prediction* about the outcome of natural selection. So, understood in this narrow sense, the Darwinian imperative would then be nothing more than a useful heuristic for interpreting in analogical terms the general condition (9) derived by the DF, recipient-centered approach. But to say this would amount, in a sense, to avoid addressing the issue of the causal representation of the fitness structure involved in the evolution of altruistic traits, that is, to avoid addressing the question of whether the expected benefits (which figure explicitly in the Darwinian imperative) play a real causal role in the evolution of altruism (like the one they play, for instance, in the evolution of reciprocal behaviours) or not at all.

The second option is quite opposed to Skyrms' view, and consists in arguing that, so far as we are interested in causal issues (and not in the mere predictive power of analogical models), Jeffrey's model is not an appropriate way of thinking about the natural selection of altruistic traits. This is the solution that I will develop in the next section. Here, I will show that the proper way of describing the genuine causal structure of altruism consists in using a modified version of Stalnaker's general formula (10), using as a maximand both expected direct *and* indirect benefits instead of mere expected reproductive success. As we have seen, Stalnaker's model (unlike Jeffrey's model) allows us to take into account the causal efficacy of the social strategies, and also embodies, for this very same reason, an excellent criterion for

identifying the causal structure associated with altruistic behaviours. From there, it will also follow that natural selection and rational deliberation, contrary to what both Skyrms (1994, 1996) and Sober (1998) argue, go "hand on hand"; for as I will show, the strategy that natural selection favours when inequalities (1) and (2) are satisfied (namely altruism) is actually the one that a rational agent (in the sense of Stalnaker's model of rationality, and in full conformity to the dominance principle) would have chosen in these circumstances.

V] A parallel between causal decision theory and IF theory

The problem I will address in this section consists in determining *which of* the indirect benefits or of the correlated (direct) benefits expected from the partners are causally responsible for the evolution of altruism, at least in those cases where the dimension of sacrifice is not *a priori* obvious. To this end, I will use Stalnaker's previous model of expected utility (instead of Jeffrey's model) as a basis for elaborating a model of expected fitness for both altruism and selfishness, so as to discriminate genuine causal influences (i.e. direct fitness effects on partner through altruistic behaviour) from any spurious correlation that may result from the social interactions between partners.

The general setting for my argument is that of a large population, structured in correlated pairs of individuals playing one-shot prisoner's dilemma, where the pairs are reformed after each interaction.[7] I will also suppose, for the sake of simplicity, that correlations are fully determined by an extraneous mechanism of limited dispersal, imposing a positive relatedness of degree r between each pair of i and j individuals.

[7] No specific assumption has to be made here according to the kind of individuals that could be represented in this setting. It may be, for instance, bacteria, but also plants, insects, or mammals.

Other kinds of assortment mechanisms (e.g. based on phenotypic recognition) will be discussed in section VI.

I will now make two further assumptions. First, I will assume that the general amount of phenotypic correlation is entirely determined by the general degree of genetic correlation, that is:

$$(13)\ P(A_j|A_i) - P(A_j|S_i) = r_{i,j}$$

Thus, the amount of phenotypic correlation that we observe among the individuals in the population is fully reflected by an identical amount of genotypic correlation. This assumption will allow us to exclude all of the complex cases where, as in some interspecific mutualisms, we observe positive phenotypic correlations without any (or less) genetic correlations. Second, I will assume that the two strategies A and S are fixed, which means that no individual can change her phenotype in the population. Thus, the mere possession of the altruistic or the selfish genotype fully determines the phenotype of the individuals.

Now, let's suppose that Hamilton's rule is satisfied in the population, so that $rb_j - c_i > 0$, $\forall i,j$ (with $i \neq j$). Automatically, the Darwinian categorical imperative is satisfied, given equality (13). Moreover, both give us the right prediction concerning the outcome of natural selection in this situation, viz. the evolution of altruism. Yet, of these two principles, I will show that only Hamilton's rule admits a legitimate causal interpretation—as only Hamilton's rule provides us with the proper representation of the causal factors at work in the evolution of altruism.

Consider first the perspective of a (randomly picked) focal altruist, I, that enters a new pair in the population with a partner, J. By adopting the perspective of this

individual, we can easily figure out the different fitness flows within the pair {I, J}. One the one hand, we have both the direct fitness cost of altruism ($-c_I$) and the indirect benefits rb_I provided by our focal altruist to her partner. Because these costs and benefits do not depend on the phenotype of the partner, their probability of occurring is equal to one—conditional of course on the focal individual effectively behaving according to her own type. On the other hand, we have the expected benefits caused by the partner. Because our focal individual is an altruist, there is a probability $P(A_J|A_I)$ that she will receive some direct benefits b_J and some indirect costs ($-rc_J$) from her partner, and a probability $P(S_J|A_I)$ that she will receive nothing at all if paired with a selfish individual. So the expectation of the direct benefits/indirect costs caused by the partner is simply equal to $P(A_J|A_I)(b_J - rc_J)$.

In this context, I will now make a simple counterfactual test and ask: what would happen if our focal altruist was "allowed" to behave selfishly *while* keeping the same (altruistic) genotype, namely, *as if* she had the exceptional possibility to choose, at this very precise moment, to behave selfishly? As we will see, the purpose of this thought experiment is merely to isolate the causal consequences of a change in the phenotype of the focal individual on the global frequency of altruism genotype, holding fixed all the values of the other variables in the population, such as the genotypic values of I and J, the coefficient of relatedness, and the conditional probabilities $P(A_J|A_I)$ and $P(A_J|S_I)$. However, in order to formulate this test in the terms of causal decision theory, we need to start by modelling the "choice" of our focal individual in terms of Stalnaker's model of expected utility (formula 10), assuming the maximizing agent analogy.

To this end, I will distinguish first between (i) the types of each member *after* the formation of the pairs (i.e. after the assortment) but *before* the interaction take place

and (ii) the effective behaviour of each member once the interaction is engaged (I consider the interaction as engaged when the focal individual has "chosen" her action/phenotype). I will denote the former by A and S, and the latter by A′ and S′. Using this notation together with Stalnaker's model, the respective (and expected) structure of costs and benefits for A′ (behaving altruistically) and S′ (behaving selfishly, which amounts to cheating here) can be represented as follows:

(14) $EW_I(A') = P(A_I' \rightarrow A_J')[rb_I - c_I + b_J - rc_J] + P(A_I' \rightarrow S_J')[rb_I - c_I]$

(15) $EW_I(S') = P(S_I' \rightarrow A_J')[b_J - rc_J] + P(S_I' \rightarrow S_J')[0]$

In these formulae, $EW_I(X')$ represents the *whole* expected cost/benefit structure associated with the strategy X′ (either A′ or S′) for the focal individual I, as it takes into account *both* the "correlated benefits/costs" caused by the partner J as well as the indirect benefits generated by the focal individual (remember that the point of our counterfactual test is precisely to assess the causal relevance of these respective benefits/costs—i.e. the "indirect" vs. the "correlated/expected" benefits/costs—with respect to the change in frequency of the altruistic genotype).

Now, a crucial assumption of the maximizing agent analogy has to be taken into account: according to this perspective, the focal individual is supposed to possess *exactly the same information as the modeller* concerning both her own genotype and the entire population structure. From a biological point of view, this assumption is justified by the fact that natural selection is, ultimately, the one which "chooses" among the types, according to their effective successes. Hence, because its "choice" is based on *all* of the relevant parameters affecting the individual's fitness (natural selection is "omniscient"), we must assume the same "knowledge" from the part of our focal individual, i.e. the actor or maximizing agent.

Transposed in the context of our previous example, this assumption merely implies that the focal individual "knows" her initial type A_I, and "knows" that the strategy of her partner is fixed by her type, A_J or S_J. Consequently, the focal individual (i.e. natural selection through the eyes of the focal individual) also "knows" that the choice of her phenotype will not affect the behaviour of her partner. Hence, being a maximizing agent (by hypothesis), she should adjust accordingly the values of the probabilities of counterfactuals according to:[8]

(16) $P(A_I' \rightarrow A_J') = P(S_I' \rightarrow A_J') = P(A_J')$

(17) $P(A_I' \rightarrow S_J') = P(S_I' \rightarrow S_J') = P(S_J')$

Besides, the focal individual should also know (as the strategies are supposed to be fixed) that the effective strategy A_J' of her partner J is fully determined by her own altruistic type A_J. So, combining this fact with what she knows about the actual pattern of assortment and about her own initial altruistic type A_I, her best estimation for $P(A_J')$ and $P(S_J')$ should be:[9]

(18) $P(A_J') = P(A_J|A_I)$

(19) $P(S_J') = P(S_J|A_I)$

Finally, in substituting these new weights in the formulae (14) and (15), and simplifying, one obtains:

(20) $EW_I(A') = P(A_J|A_I)(b_J - rc_J) + rb_I - c_I$

(21) $EW_I(S') = P(A_J|A_I)(b_J - rc_J)$

[8] It is important, however, to note that (16) and (17) depend entirely of the biological setting considered. For instance, one could well have imagined other biological situations where our focal individual would have had the possibility of *forcing* her partner to behave altruistically toward him, so that these equalities would not have been valid.

[9] Given that the actor knows that $P(A'_J) = P(A_J)$ (i.e. that the strategies are fixed), one has indeed: $P(A_J') = P(A_J|A_I)P(A_I) + P(A_J|S_I)P(S_I)$. In addition, the actor knows her own type A_I with certainty, so that $P(A_I) = 1$. Hence: $P(A_J') = P(A_J|A_I)$. The same kind of reasoning holds for determining $P(S_J')$ in equality (17).

From this, one then recovers the classic form of Hamilton's rule by subtracting (21) from (20): $rb_I - c_I$.

Comparing the formulae (20) and (21), two general remarks can be made. First, a switch from A′ to S′ does not affect the expected benefits/costs $P(A_J|A_I)(b_J - rc_J)$ caused by the partner and received by the focal individual. This is because these benefits/costs only depend on the behavioural type of the partner, which is *already* determined by the mechanism of assortment responsible for the population structure. Thus, what this counterfactual test first reveals is that the correlation measured by the conditional probability $P(A_J|A_I)$ is a *spurious* one—at least with regard to our purpose, which consists in explaining how a focal individual could increase the representation of her type in the global population. Hence, it cannot be used in order to substantiate a causal interpretation of the correlated benefits, that is, to substantiate the idea that, in behaving altruistically, the actor would indirectly cause her partner to provide her with some expected (direct) benefits.

Second, in choosing to behave selfishly, our focal individual would not incur the cost of altruism. But at the same time, the indirect pathway for genes' flow rb_I—which is actually the only thing she controls (besides $– c_I$)—would also be turned off, for there would be no more indirect benefits generated by the actor in this scenario. So the gene's frequency of altruism would have slightly decreased at the end of the interaction (all things being equal).

Thus, from that simple counterfactual test, we can conclude that the expected fitness benefits/costs $P(A_J|A_I)(b_J - rc_J)$ are *causally irrelevant* to the evolution of altruism. Indeed, what the actor "chooses" ($A_I′$ or $S_I′$) does not have any effect on her partner's behaviour, but only has an effect on her partner's reproductive success (through rb_I). So, from the perspective of a focal individual seeking to increase the

representation of her type, these expected benefits/costs are not an "incentive" to behave selfishly.

As this point, it is now clear that there is a tight parallel between the way IF theory deals with the evolution of altruism and the way causal decision theory deals with Newcomb-like problems. As an actor-centered perspective, the IF approach has indeed an intrinsic sensitiveness to spurious correlations, because it always considers assortment *as a given* in order to focus precisely on the sole causal consequences of a particular change in the phenotype of the focal individual on the relevant components of fitness (West and Gardner 2013; Birch 2016). In this respect, both causal decision theory and IF theory have in common their most fundamental characteristic, namely their focus on the *causal efficacy* of the strategies, regardless of their (prima facie) auspiciousness.[10] By contrast, the Darwinian categorical imperative (born from the transfer of Jeffrey's model of expected utility in the evolutionary setting) does not provide any insight on the causal structure of the process at work in the evolution of altruism, even though it provides a correct prediction concerning the outcome of this process.

In conclusion to this section, it must be noted that the previous counterfactual test has also strong affinities with the famous "mutation test" designed by Nunney (1985). This test, initially, was designed to discriminate weak (or "false") altruism—which evolves through the sole direct pathway—from true altruism (i.e. strong altruism)—which can't evolve without indirect benefits. The method is quite simple, and consists in asking what would happen to the absolute (direct) fitness of a (prima facie) altruistic

[10] There is, of course, a noticeable difference between the two approaches (IF theory and causal decision theory) concerning their appreciation of the best strategy in the correlated prisoner's dilemma; for while the causal decision theory claims that selfishness is always the rational option to choose, IF theory, in contrast, leads to the recognition of altruism as the optimal strategy, provided that Hamilton's rule is satisfied. However, this discrepancy is inessential, as it merely stems from the contingent (and obvious) fact that the relevant maximand is not the same for social evolutionary theory and for classical rational decision theory—in the latter, the agents only care about their own personal utility.

individual if she suddenly *mutated* into a selfish individual, keeping constant her social environment. From this thought experiment, Nunney (1985) concluded that if the focal mutant had her direct fitness maximized, this would mean that the trait is not an altruistic trait, and evolves like any other traits under the action of individual selection. If instead she had her direct fitness decreased, then the only possible explanation for why such a trait evolves would be to invoke kin selection, and to interpret this trait as a true instance of altruism.[11]

There is, however, a crucial difference between Nunney's mutation test and the argument developed here; for contrary to the mutation test, the genotype of the focal individual is held fixed in our counterfactual test, and only her phenotype is allowed to change. The reason for holding fixed her (geno)type is that, without this assumption, there would be no more indirect benefits within the pair; for once we cease to keep fixed the value of the genotype of the actor by introducing a mutation event, the value of relatedness falls to zero within the pair, and altruism is no longer the optimal strategy from the viewpoint of the focal actor. Thus, because my point was primarily to outline the causal role of the indirect benefits in the evolution of altruism, I had to integrate this condition as a part of my argument.

VI] An objection: altruists control assortment

I would now turn to an important objection that can be raised against the argument developed in the previous section. This objection, in a nutshell, amounts to the claim that the kind of counterfactual analysis developed above overlooks the biological fact that altruist individuals *control* the pattern of assortment (Rosas 2010).

[11] Again, I leave deliberately open the familiar issue of the relationship between kin selection and group selection here, which is not directly relevant for my argument.

As we have seen, an essential condition of the analogy between causal decision theory and IF theory is that, whatever the focal individual "decides" to do, the behavioural type of her social partner(s)—i.e. her social environment—has to be considered as a fixed parameter resulting from the mechanism of assortment. In the previous section, this clause was already built in the model, as the pairs were supposed to form according to an extraneous (and unspecified) limited dispersal mechanism, with a degree r of relatedness between individuals. But admittedly, one could legitimately wonder about its biological justification. Most of the time, indeed, the patterns of assortment are not "imposed" on the members of the population *from the outside*, like in our previous model, but appear to be emergent properties of biological populations, determined by the phenotypes of the individual organisms. The most obvious instance of such emergent relations is probably when social interactions between individuals are determined by some specific mechanisms of *phenotypic recognition* that are themselves an integral part of their individual phenotype. But even limited modes of dispersal are usually constrained by some basic phenotypic properties of organisms, like their mobility.

From these observations, it thus appears that altruists usually control assortment. But shouldn't this also imply the following conclusion, namely that to *switch* from an altruistic to a selfish phenotype would lead—most of the time—to the exclusion of the focal individual from the altruistic groups, *thus decreasing her DF*?

This argument, which has been developed by Rosas (2010) against the logic of the IF approach, deserve a closer look. According to Rosas, the decision of holding constant the social environment—i.e. the phenotype of the partner(s)—always happens by *fiat* in the IF reasoning, because it does not take into account the fact that altruism and assortment are ultimately two inseparable components of a *single* causal

evolutionary process, which leads to the maximization of the DF of the altruists. Rosas' argument can be reformulated as follows. If one accepts the proposition that assortment with phenotypically similar partners is ultimately conditional on the fact of being of an altruist type, then ceasing to be an altruist would have the direct effect of excluding the focal individual from the interactions with other altruists. But such an exclusion would automatically result in a marginal decrease of the DF of this new selfishly behaved individual, thereby contradicting the view of altruism as a self-sacrificing trait constitutive of the IF theory. From this reasoning, Rosas concludes: "altruists that have stably evolved will have done so because they control assortment; and in this sense the benefits they receive are always due to their own traits or action." (Rosas 2010, 7). This, he claims, is "the fundamental conceptual insight underlying the DF perspective." (ibid.).

As I will now show, however, this whole reasoning is flawed in several respects. A first objection that one can address to Rosas' argument is that its validity relies on a misunderstanding concerning the meaning and the implication of its first premise, namely the proposition that "altruists control assortment". There is no doubt that assortment mechanisms, as Rosas rightly emphasizes, are very often *endogenous* characters of individual altruists, and that the two coevolve in the process of social evolution so as to produce complex altruistic adaptations. But this undeniable empirical fact, contra Rosas, does not *logically* entail that a given local change in the altruistic phenotype of a focal individual would always (or at least in most of the situations) *cause* a change in the composition of her social environment.

Consider for instance the well-known thought experiment of the "greenbeard effect" (Dawkins 1976). In this case, altruism is supposed to be a pleiotropic effect of an allele coding for a phenotypic marker—the greenbeard—that also determines

assortment between altruists, so that the expression of altruism is supposed to be both accompanied by and conditional on the greenbeard phenotypic marker. However, it is clear in this context that both altruism and the marker which determines its actualization are simply joint effects of a common genetic cause, so that a given change in the altruistic phenotype of a given focal individual, holding all parameters constant (including her genotype), would *not* change anything to the actual pattern of assortment. Cheating is thus perfectly possible in such a situation, and a variant of the previous counterfactual test could thus as well be used in that kind of situation to show that the evolution of greenbeard genes is ultimately caused by the indirect benefits provided by the actors to the other carriers of the same marker.

An important remark, however, has to be made here concerning the application of the IF maximizing agent analogy to greenbeard-like cases. Unlike what happens when the value of r is fixed by kinship patterns, cheating is usually the strategy favoured by natural selection in those cases (West & Gardner 2009). At first, this fact seems to contradict our argument. But it has, actually, a simple explanation in IF fitness terms. In the greenbeard thought experiment, green-bearded individuals are supposed to be perfectly related at the locus for altruism, so that $r = 1$ and $b - c > 0$. Thus, as long as we exclude a priori the opportunity of cheating, altruism is the strategy that maximizes the genotypic contribution of the actors to the future generations via its indirect benefits. However, if we allow for the possibility of cheating, that is, of individuals who would keep their greenbeard while ceasing to behave altruistically (such individuals are called "falsebeards" in the literature), the value of r at the locus for altruism *falls to zero* between green-bearded individuals.

The explanation to this fact is that the value of r, in this case, is a direct function of the reliability of the greenbeard as a signal of the presence of an altruistic genotype

(Frank 1998): indeed, when the greenbeards (i.e. the phenotypic markers) cease to be correlated with the genotype of their bearers at the locus for altruism, the probability of these same bearers being true altruists—from the viewpoint of a focal actor—is simply equal to the global frequency of altruists in the subpopulation of green-bearded individuals; hence $\forall i,j\ P(A_j|A_i) - P(A_j|S_i) = 0$. Thus, one can see that when cheating is included as an available option in the strategy set, the DF benefits become the main *causal* determinants of the phenotypic change driven by natural selection; and because the cost of altruism is no longer compensated within the green-bearded subpopulation by any indirect benefits, altruism is no longer an optimal strategy from the viewpoint of an IF maximizing actor.

This result, however, is perfectly consistent with the kind of IF version of Stalnaker's model proposed in section V. Of course, one could retort that there are some cases where, contrary to greenbeard-like cases, altruism and the phenotypic markers are effectively linked *in non-spurious ways*—the limiting cases being the ones in which the phenotypic markers constitute themselves an honest signaling of altruist condition. For instance, one could invoke the cases where altruism (the honest signal *per excellence* of altruist condition) is itself the marker on which assortment is conditioned. For then, ceasing to behave altruistically would have the consequence of excluding oneself from the apparent "win-win" structure of the altruists' interactions. But even in those cases (i.e. assuming that altruism causes assortment, properly speaking), my claim is that Rosas' conclusion is not justified. Here again, one has to distinguish carefully between the cases where cheating is allowed from the cases where cheating is not a likely biological possibility.

First, Rosas' argument does not apply if we allow for the *possibility of cheating*, that is, of individuals which could *fake* systematically the altruistic trait at the expense

of their partner—thus breaking the rule of assortment. For in those cases, behaving altruistically would no longer be the cause of the interaction of the other altruists, as it would no longer guarantee a higher probability of interacting with an altruist than with a cheater (regardless of their frequency). Hence, Rosas' premise would not be satisfied, and by using our counterfactual test (i.e. by keeping constant the phenotype of the partner), one would find in that case that altruism only evolves thanks to indirect benefits.

Now, Rosas is right to stress that the control of assortment by altruists makes a difference to the *stability* of altruism at the population level. In particular, it is reasonable to assume that the more altruists can control assortment (e.g. due to the evolution of cognitive abilities to detect cheaters through phenotypic clues), the more altruism will be stable in biological populations—though other independent factors such as kinship also play a crucial role from this respect (Okasha 2002). But the issue of stability must not be conflated with that of identifying the causal fitness components involved in the evolution of altruism. For as long as cheating is a likely opportunity, the application of the counterfactual test will reveal that indirect benefits play the central role in the evolution of altruism.

Second, when cheating is not a likely biological possibility, i.e. when one cannot *plausibly* imagine a particular biological situation where altruism could be dissociated from the marker which actually determines the assortment, *then* the trait considered is simply *not* an altruistic one. For this implies that behaving selfishly involves a net DF cost to the individual.

This point is not actually a new one. Indeed, Ridley and Grafen (1981) already noticed for the greenbeard case that if the phenotypic association between altruism and the greenbeard marker were not plausibly defeasible by a modifier gene, then the sort

of altruism induced by the greenbeard allele could not be considered as a harm on the whole genome—whatever the value of relatedness at the other loci—and so could not properly be considered as a form of true altruism. Put simply: in those cases where the opportunity for cheating is a priori excluded, there is no necessity for Rosas' conclusion to follow either (for even there, it is wrong to claim that altruism evolves thanks to *direct* benefits). One shall note, however, that the issue of deciding the strength of such an association remains ultimately an empirical one; for it could well appear that a trait which was previously considered as non-altruistic turns out to be an altruistic one after further analysis.

VII] Conclusion

Most of the time, the evolutionists who use the DF approach envisage it as a more convenient way of deriving Hamilton's rule, and the IF approach as a more natural way of interpreting the terms of the rule, while arguing for their conceptual equivalence. But this consensus, as I have shown, is misleading.

IF theory, indeed, does not merely provide the modeller with a more intuitive heuristic in order to make sense of the partition of cost and benefits of altruism that one obtains from the DF approach. But as it turns out from the analogy between this approach and causal decision theory (which claims that a strategy can only be deemed a rational one by a focal agent if it turns out to have any sufficient causal efficacy to bring the expected outcome), the IF approach has actually an exclusive sensitiveness to the causal consequences of altruistic traits—contrary to the DF which, envisaged as a description of the genuine fitness structure, may lead in some circumstances to conflate spurious correlations with genuine causal relations, and to overlook the

dimension of sacrifice proper to strong altruism. Thus, whatever the practical advantages of the DF approaches, and despite its mathematical equivalence with the IF approach, the latter remains the only proper way of describing the causal structure of altruistic traits.

References

Birch, J. 2016. "Hamilton's Two Conceptions of Social Fitness" *Philosophy of Science,* 83: 848-860.

Bergstrom, Theodore C. 2002. "Evolution of Social Behaviour: Individual and Group Selection." *Journal of the Economic Perspectives* 16(2): 67-88.

Cartwright, Nancy. 1979. "Causal Laws and Effective Strategies." *Noûs* 13: 419-437.

Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Fletcher, Jeff A., and Michael Doebeli. 2009. "A Simple and General Explanation for the Evolution of Altruism." *Proceedings of the Royal Society. Series B. Biological Sciences* 276: 13-19.

Frank, Steven A. 1998. *The Foundations of Social Evolution*. Princeton: Princeton University Press.

Gardner, Andy, and Stuart A. West. 2009. "Greenbeards." *Evolution* 64: 25-38.

Gardner, Andy, Stuart A. West and Geoff Wild. 2011. "The Genetical Theory of Kin Selection", *Journal of Evolutionary Biology* 24 (5): 1020-1043.

Gibbard, Allan, and William Harper. 1981. “Counterfactuals and Two Kinds of Expected Utility.” In *IFS: Conditionals, Beliefs, Decision, Chance, and Time*, eds. William Harper, Robert Stalnaker, and Glenn Pearce, 153-190. Dordrecht: Reidel.

Grafen, Alan. 1999. “Formal Darwinism, the Individual-as-maximising-agent Analogy, and Bet-hedging. *Proceedings of the Royal Society. Series B. Biological Sciences* 266: 799-803.

Grafen, Alan. 2006. “Optimization of Inclusive Fitness.” *Journal of Theoretical Biology* 238: 541-563.

Hamilton, William D. 1964. “The Genetical Evolution of Social Behaviour I, II.” *Journal of Theoretical Biology* 7 (1): 1-16, 17-52.

Hamilton, William D. 1970. “Selfish and Spiteful Behaviour in an Evolutionary Model.” *Nature* 228: 1218-1220.

Hamilton, William D. 1975. “Innate Social Aptitudes of Man: an Approach from Evolutionary Genetics.” In *Biosocial Anthropology*, ed. Robin Fox, 133-155. New York: Wiley.

Jeffrey, Richard. 1965/1983. *The Logic of Decision*. Repr. Chicago: University of Chicago Press.

Joyce, J. (1999). *The Foundations of Causal Decision Theory*, Cambridge: Cambridge University Press.

Kerr, Benjamin, Peter Godfrey-Smith and Marcus K. Feldman. 2004. “What is Altruism?” *Trends in Ecology and Evolution* 19 (3): 135-140.

Lehmann, Laurent, and Laurent Keller. 2006. “The Evolution of Cooperation and Altruism: a General Framework and Classification of Models.” *Journal of Evolutionary Biology* 19: 1365-1376.

Leigh, Egbert G. 2010. “The Group Selection Controversy.” *Journal of Evolutionary Biology* 23: 6-19.

Lewis, David. 1979. “Prisoner's Dilemma is a Newcomb Problem.” *Philosophy and Public Affairs* 8: 235-240.

Lewis, David. 1981. “Causal Decision Theory.” *Australasian Journal of Philosophy* 58: 5-30.

Marshall, J.A.R. 2011. “Group selection and kin selection: formally equivalent approaches.” *Trends in Ecology and Evolution* 26: 325-332.

Martens, J. 2017. “Inclusive fitness and the maximizing agent analogy”, *British Journal for the Philosophy of Science*; 68: 875-905.

Maynard Smith, John. 1964. “Group Selection and Kin Selection.” *Nature* 201: 1145-1147.

Maynard Smith, John. 1982. *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Nozick, Robert. 1969. “Newcomb's Problem and Two Principles of Choice.” In *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher, 114-146. Dordrecht: Reidel.

Nunney, Len. 1985. “Group Selection, Altruism, and Structured-deme Models.” *American Naturalist* 126(4):212-230

Okasha, Samir. 2002. “Genetic Relatedness and the Evolution of Altruism.” *Philosophy of Science* 69: 138-149.

Okasha, S. (2016). “On Hamilton’s Rule and Inclusive Fitness Theory with Non-additive Payoffs”. *Philosophy of Science,* 83: 873-883.

Okasha, Samir, and Johannes Martens. 2016. “The causal meaning of Hamilton’s rule”. *Royal Society Open Science* 3(3).

Ridley, Mark, and Alan Grafen. 1981. “Are Green Beard Genes Outlaws?” *Animal Behaviour* 29: 954-955.

Rosas, Alejandro. 2010. “Beyond Inclusive Fitness? On a Simple and General Explanation for the Evolution of Altruism” *Philosophy and Theory in Biology* 2: 1-9.

Savage, Leonard. 1954. *The Foundations of Statistics*. New York: Wiley.

Skyrms, Brian. 1980. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven, CT: Yale University Press.

Skyrms, Brian. 1994. “Darwin Meets the Logic of Decision: Correlation in Evolutionary Game Theory.” *Philosophy of Science* 61: 503-528.

Sober, Elliott. 1998. “Three Differences Between Evolution and Deliberation.” In *Modeling rationality, morality and evolution,* ed. Peter Danielson, 408-422, Oxford: Oxford University Press.

Stalnaker, Robert. 1972. “Letter to David Lewis.” In *IFS: Conditionals, Beliefs, Decision, Chance, and Time,* eds. William Harper, Robert Stalnaker, and Glenn Pearce (1981), 151–152. Dordrecht: Reidel.

Taylor, Peter D., and Steven A. Frank. 1996. “How to Make a Kin Selection Model.” *Journal of Evolutionary Biology* 180: 27-37.

Taylor, Peter D., Geoff Wild, and Andy Gardner. 2007. “Direct Fitness or Inclusive Fitness: How Should We Model Kin Selection?” *Journal of Evolutionary Biology* 20: 301-309.

Trivers, Robert L. 1971. “The evolution of reciprocal altruism.” *Quarterly Review of Biology* 46: 35-57.

von Neumann, John, and Oskar Morgenstern. 1947. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

West, Stuart A., Ashleigh S. Griffin, and Andy Gardner. 2007. “Social semantics: Altruism, Cooperation, Mutualism, Strong Reciprocity and Group Selection”. *Journal of Evolutionary Biology* 20: 415-432.

West, Stuart A. and Andy Gardner. 2013. “Adaptation and Inclusive Fitness”, *Current Biology*, 23, pp. R577–84.

Wilson, David Sloan. 1979. “Structured Demes and Trait-group Variation.” *American Naturalist* 113: 606-610.