Biol Philos (2011) 26:637–654
DOI 10.1007/s10539-011-9273-3

Evolutionary game theory, interpersonal comparisons and natural selection: a dilemma

Till Grüne-Yanoff

Published online: 10 June 2011

Abstract When social scientists began employing evolutionary game theory (EGT) in their disciplines, the question arose what the appropriate interpretation of the formal EGT framework would be. Social scientists have given different answer, of which I distinguish three basic kinds. I then proceed to uncover the conceptual tension between the formal framework of EGT, its application in the social sciences, and these three interpretations. First, I argue that EGT under the biological interpretation has a limited application in the social sciences, chiefly because strategy replication often cannot be sensibly interpreted as strategy bearer reproduction in this domain. Second, I show that alternative replication mechanisms imply interpersonal comparability of strategy payoffs. Giving a meaningful interpretation to such comparisons is not an easy task for many social situations, and thus limits the applicability of EGT in this domain. Third, I argue that giving a new interpretation both to strategy replication and selection solves the issue of interpersonal comparability, but at the costs of making the new interpretation incompatible with natural selection interpretations of EGT. To the extent that social scientists seek such a natural selection interpretation, they face a dilemma: either face the challenge that interpersonal comparisons pose, or give up on the natural selection interpretation. By identifying these tensions, my analysis pleas for greater awareness of the specific purposes of EGT modelling in the social sciences, and for greater sensitivity to the underlying microstructure on which the evolutionary dynamics and other EGT solution concepts supervene.

Keywords Evolution · Gene theory · Selection · Social science · Economics · Interpersonal utility comparisons

T. Grüne-Yanoff (✉)
Helsinki Collegium for Advanced Studies, University of Helsinki,
P.O. Box 4, FI-00014 Helsinki, Finland
e-mail: till.grune@helsinki.fi

Introduction

In the last two decades, social scientists, in particular economists, have increasingly employed evolutionary game theory (EGT) in their disciplines. Yet the formal framework of EGT had been developed by biologists, with a firm focus on a population genetics interpretation. So the question arose how and to what extent this framework required re-interpretation to be applicable in the social sciences.

Social scientists have differed in their answer to this question. In this paper, I distinguish three basic kinds of responses. First, some social scientists accepted all main tenets of the biologists' interpretation. Second, some others replaced the biologists' interpretation of strategy replication. Third, some replaced both the biologists' interpretation of strategy replication and of strategy selection.

The purpose of this paper is to uncover the conceptual tension between the formal framework of EGT and its application in the social sciences under these three interpretations. First, I argue that EGT under the biological interpretation has a limited application in the social sciences, chiefly because strategy replication often cannot be sensibly interpreted as strategy bearer reproduction in this domain. Second, I show that those social scientists who replace the problematic reproduction interpretation with alternative replication mechanisms face the challenge that all of these mechanisms imply the interpersonal comparability of strategy payoffs. Giving a meaningful interpretation to such comparisons is not an easy task for many situations in the social sciences, and thus limits the applicability of EGT in this domain. Third, I argue that giving a new interpretation both to strategy replication and selection solves the issue of interpersonal comparability, but at the costs of making the new interpretation incompatible with natural selection interpretations of EGT. To the extent that social scientists seek such a natural selection interpretation, they thus face a dilemma: either face the challenge that interpersonal comparisons pose, or give up on the natural selection interpretation. For some situations, one horn of this dilemma may be more acceptable than the other, but both often remain disadvantageous. Between these two horns, the use of EGT to the social sciences is significantly diminished.

The paper is structured as follows. "The formal basics of EGT" describes the basics of the formal EGT framework. "The biological interpretation" sketches EGT's biological interpretation, and its requirement of interpersonal payoff comparisons. "Transfer into the social sciences" describes the main motives of importing EGT into the social sciences. It also discusses some of the main problems with adopting the biological interpretation unchanged. "A new interpretation of strategy replication" presents efforts to replace the biological interpretation of strategy replication with alternative mechanisms, at the hand of the most prominent models of imitation and experimentation. I show that interpersonal payoff comparisons are required, not only for these specific models, but generally for models that seek to only re-interpret strategy replication and nothing else. "Digression: are inter-personal payoff comparisons problematic?" briefly discusses why such interpersonal comparisons pose a challenge for many social science applications. "A new interpretation of strategy selection" presents efforts to replace the biological interpretation both of strategy replication and of strategy selection,

and discusses some disadvantages of these re-interpretations. "EGT and natural selection" argues that the discussed re-interpretations of strategy selection are incompatible with natural selection interpretations of EGT.

The formal basics of EGT

EGT came into existence between 1972 and 1982 through the employment of the mathematical theory of games in biological contexts. This application gave rise to a novel formal framework, which was still recognisably game-theoretic, but contained its own representational tools and solution concepts. When from the 1980s onwards social scientists adopted EGT for their own purposes, the most visible sign of the transfer from biology was the import of this framework into their disciplines. The formal framework thus is the most obvious element that connects EGT in biology and in the social sciences. In this section, I give a brief overview of it, while the rest of the paper will be concerned with its different interpretations in the respective disciplines.

Today, EGT is too broad and diverse a field to have its formal framework comprehensively overviewed in a research article. However, by tracing the history of EGT, and by surveying its different contemporary uses, one can identify its core idea, its main forms of analysis, and its most dominant solution concepts.[1] This will suffice for the purpose of this paper.

The core idea of EGT is that in interactive situations, strategy replication depends on strategy performance in the given population. Strategy performance, as in classical game theory, is represented as payoffs assigned to all strategy profiles.[2] However, in contrast to classical game theory, EGT focuses not on decisions of individual players, but on properties of the whole population, and on the effect of properties of previous populations on future population.

Two main forms of analysis can be distinguished. EGT provides a *static* analysis, essentially of stability properties of populations, and a *dynamic* analysis, of how strategy frequencies change in a population over time. All uses of EGT in biology and the social sciences fall into one of these categories.

The most prominent solution concept employed in the static analysis is the Evolutionary Stable Strategy (ESS). With the help of ESS, EGT investigates whether a certain frequency distribution can resist the introduction of (a small number of) additional strategies from a specified list of strategy types. To give a very simple example, one may ask what happens when a single y-type strategy is introduced into a population consisting only of x-type strategies. If the y-type strategy does better than the x-type strategies in that situation, then y will replicate at a higher rate than the xs, so that the initial frequency distribution (i.e. xs only) will not be restored. We then say that x is not an ESS, because it cannot resist invasion

[1] For a historical investigation, see Grüne-Yanoff (2011).

[2] A strategy profile is a combination of strategies $s = \{x_1,\ldots,x_n\}$ for each of the n players, which fully specifies all actions in a game. A well-defined game has payoffs assigned to every strategy profile. I will often speak of the payoffs of a strategy x_i—by which I mean the set of payoffs assigned to x_i for each strategy profile s^{-i}.

by y. If however the y-type strategy does worse than the x-type strategies in that situation, then the xs will replicate at a higher rate than the y, and eventually the initial frequency distribution of only xs will be restored. In that case, we say that x is an ESS, because it resists invasion by y.

Formally, a strategy x from the strategy set Σ is an ESS if and only if: (1)

1. for all $y \in \Sigma$, $u(x, x) \geq u(y, x)$
2. for $y \in \Sigma$, with $y \neq x$, if $u(x, x) = u(y, x)$, then $u(x, y) > u(y, y)$ (Weibull 1995, 37)

That is, x is an ESS either if x obtains a higher payoff, when played against itself, than any other strategy when played against x; or if there is another strategy y that obtains the same payoff when played against x as x obtains when played against itself, then x does better when played against y than y does when played against itself.

To determine the ESS, it is required both to compare the payoff of strategy x with payoffs of other strategies, as well as to compare the payoffs of two strategies x when played against different strategies. Thus the notion of ESS clearly requires *inter-strategy comparisons* of payoffs. While this implication is formally innocuous, it forms the basis for the later discussion of different interpretations of the formal framework.[3]

The most prominent approach employed in the dynamic analysis is the replicator dynamics (RD). The RD describes the change in the population share of each strategy in time.[4] More specifically, the continuous RD is derived in the following way. A population state is defined as the vector $x(t) = (x_1(t),\ldots, x_k(t))$, where each component $x_i(t)$ is the population share of strategy i at time t. The expected payoff to any pure strategy i in a random match, when the population is in state x, is accordingly $u(e^i, x)$. The associated population average payoff is $u(x, x) = \Sigma_i x_i{*}u(e^i, x)$. Now this is coupled with the idea that a strategy proliferates *in linear proportion* to the difference between its payoff and the population average payoff:[5]

$$dx_i/dt = \left[u\left(e^i, x\right) - u(x, x)\right] * x_i \quad (2)$$

(Weibull 1995, 72).

[3] Alternative stability concepts used for static analysis—for example neutral stability (Weibull 1995, 46) or robustness against equilibrium entrants (Weibull 1995, 48)—differ in requiring that no other strategy earns a higher payoff than an incumbent, or in restrictions on the kind of other strategies; but all require inter-strategy comparisons in the same way as ESS.

[4] There are both continuous and discrete versions of the RD, of which the continuous variant is the more prominent. The latter is discussed here.

[5] The relation between proliferation and payoffs characterizes different classes of selection dynamics. While a linear relation characterizes the RD, wider classes are characterized by payoff positivity and payoff monotonicity, respectively (Weibull 1995, 139–152). The RD is by far the most prominent selection dynamic in EGT, and discussion therefore focuses on it.

That is, the change in strategy's x_i's population share in time t is determined by x_i's current population share and the difference between its expected payoff and the population average payoff.

It is clear from this definition that payoffs of different strategies are aggregated and compared. Hence the replicator dynamic requires the comparison of payoffs across strategies. More specifically, it requires the possibility of comparing payoffs on an interval scale, so that the sizes of payoff differences are comparable.[6]

The inter-strategy payoff comparison does not formally differentiate EGT from classical game theory. Classical game theory compares the payoffs of different strategy profiles (to the same individual), and it also requires the comparison on an interval scale to construct players' mixed strategies. However, EGT's formal framework is interpreted in a way that necessitates not only inter-strategy, but also inter-*personal* comparisons of payoffs, both in biology and in the social sciences. I will begin by sketching the biological interpretation.

The biological interpretation

Population geneticists within the discipline of biology were the first to develop EGT. It is important not to neglect these historical roots. They determined the first conceptual interpretation of the formal framework.[7] This interpretation had important effects on social scientists, convincing them that EGT was a useful theory for their purposes, and shaping the ways how they applied it.

Three aspects of population geneticists' interpretation stand out. First, they associated each strategy with an individual strategy-bearer or player. Thus, as in classical game theory, EGT models players; yet in contrast to classical game theory, these players do not have a choice of strategies, but instead are programmed to play only one strategy. Under the biological interpretation, EGT investigates how strategic interaction in a population affects the frequencies of strategy-bearer types (i.e. players programmed to play a certain strategy) in that population.

Secondly, biologists interpreted the interaction between players as *material play*. While classical game theorists model the mental deliberation process that a player undergoes, biologists used game theory to model actual interactions in biological populations, and to determine the actual outcome of play. The contrast is between modelling mental representations, like beliefs and evaluations, and modelling material events, like what players do, and which consequences obtain. Biological EGT models what strategy would be selected, given initial material conditions.

[6] The class of selection dynamics is not properly described and is expanding as I write. Yet all *payoff-based* dynamics seek to capture the relation of payoffs and strategy replication. While many of the dynamics belonging to this class can be distinguished from the RD described here by the ways how they treat this relation, and the representational frameworks they use, they *all* share with RD the requirement of inter-strategy comparisons of payoffs. Roughly speaking this requirement is necessitated by the need of relating payoffs to *relative* strategy replication, thus putting the payoffs of different strategies in a relation to each other and hence comparing them.

[7] Indeed it would be historically more accurate to say that their theoretical concepts influenced the construction of the formal framework (cf. Grüne-Yanoff 2011).

Classical game theory models what players would *believe* to be the equilibrium strategy profile, given their initial beliefs about strategies, payoffs and information sets.

Thirdly, biologists interpreted strategy replication as strategy-bearer reproduction. The mechanism that they posited starts from material interaction between players with a fixed strategy. Depending on the opponents' strategies, an interaction yields a certain consequence for the player, represented by her strategy's payoff. The payoff, interpreted as individual fitness, determines the number of offspring the player will have in the next round. Assuming that players breed true, the payoff also determines how many times the player's strategy will be replicated in the next round.

Under this interpretation, EGT not only requires inter-strategy payoff comparisons, but inter-player payoff comparisons. As each player has only one strategy, the performance of its strategy is compared with the performance of other players' strategies in the population. This is true both for the computation of the ESS and the replicator dynamics.

To see this more clearly, I write $u_x(y)$ for the payoff that a player programmed to play x obtains when playing against a player programmed to play y. The definition 1 turns into the following:

A strategy x from the strategy set Σ is an ESS if and only if: (1a)

1. for all $y \in \Sigma$, $u_x(x) \geq u_y(x)$
2. for $y \in \Sigma$, with $y \neq x$, if $u_x(x) = u_y(x)$, then $u_x(y) > u_y(y)$

It is clear from this re-description that both conditions require comparing the payoffs of different players. Thus the notion of ESS clearly requires inter-player comparisons of payoffs.

Similar for the replicator dynamics: computing the population average payoff implies summing over the payoffs of different players: $\hat{u}(x) = x_1 * u_1(x) + \cdots + x_k * u_k(x)$. This average is then subtracted from an individual player's payoff:

$$dx_i/dt = [u_i(x) - \hat{u}(x)] * x_i \quad (2a)$$

Thus, the replicator dynamic requires the comparison of payoffs across players. More specifically, it requires the possibility of comparing payoffs on an interval scale, so that the sizes of payoff differences are comparable.

This conclusion should not be surprising, as biologists sought to model a special form of natural selection—frequency-dependent selection—with EGT. Natural selection requires that some genotypes have a greater ability to survive and reproduce in comparison to other genotypes. Without such a notion of relative success, there is no basis on which selection could take place. Fitness is an exemplar case of a *relational* concept, although it is often mistaken for an intrinsic property of an individual organism (Wimsatt 1980). Modelling natural selection with EGT therefore required comparison between the phenotypically relevant consequences of different strategies, carried by different individuals. Hence payoffs must be inter-individually comparable.

Furthermore, the *explananda* of EGT are changes in frequency over subsequent generations. The frequency of a genotype in a population cannot be the property of any individual, but only that of the population. Yet EGT models individual payoffs as influencing the population frequency property: the difference between individual and average strategy payoff determines how much the frequency of that strategy will change relative to others in the next generation. In other words, payoffs represent the *relative strength* of the causal factor operating on the population frequency (Millstein 2006, 634). For this to be meaningful, individual payoffs must be comparable to each other with respect to the magnitude of their influence. Hence, again, payoffs must be inter-individually comparable.

Both the notion of relative success and of relative causal strength are captured in the notion of Darwinian individual fitness (Rosenberg and Bouchard 2010). When interpreting game payoffs as fitness, the kinds of inter-personal payoff comparisons required by biological EGT are thus covered. Interestingly, biologists have argued that the interpretation of strategy payoffs as fitness constitutes an improvement over the standard utility interpretation (Maynard Smith 1982, vii). Whatever else one may think of this claim, it is true that the fitness interpretation is free from the problems that the utility interpretation encounters with such inter-player comparisons. Biologists do not and need not worry about inter-player payoff comparisons in EGT.

Transfer into the social sciences

Things are not that simple in the social sciences. Recently, Kuhn (2004) pointed out that the apparent need to make inter-player comparisons in EGT poses a potential problem for its adoption into the social sciences.[8] In the remainder of this paper, I investigate this claim further and situate it in an analysis of three different interpretations of the EGT framework in the social sciences.

The transfer of EGT into the social sciences began about three decades ago and is still continuing. Economists in particular have found it to be an important tool in three domains of their discipline: equilibrium selection, solution concept justification and population dynamics modelling. In the 1980s, the dominant refinement program of equilibrium selection suffered from a fragmentation of the underlying rationality models, the details of which often were relevant for the ensuing selection. EGT was then seen as a promising alternative, because it seemed to offer a unifying selection mechanism that avoided this fragmentation and detail-dependence. Similar with the justification of equilibrium concepts. Many applied economists today seem to believe that their use of the Nash equilibrium is warranted by EGT (Sugden 2001). They interpret the formal convergence results of EGT as

[8] 'The evolutionary dynamics and the conditions on payoffs characterizing equilibria discussed above all seem to require that payoffs be interpersonally comparable… When our concern is biology, and payoffs are just measures of reproductive success, this is quite appropriate. When our concern is culture, however, we should be a little more wary. It is common to regard the payoffs of these games as utilities, and the questions of whether interpersonal utility comparisons are meaningful and measurable are notoriously vexed' (Kuhn 2004, 13).

showing that Nash equilibria are reached through selection processes, without any need for problematic epistemic prerequisites to be satisfied (e.g. Gintis 2000, 201). Finally, economists today use EGT to explain what they identify as selection processes in the social realm.

Three kinds of re-interpretations accompanied these transfers into the social sciences. Some social scientists accepted most of the biologists' interpretation of the EGT framework, including their interpretation of strategy replication and strategy selection. Others replaced the biologists' interpretation of strategy replication; and yet others replaced both the interpretation of strategy replication and strategy selection.

As for the first category, many economists not only adopted the formal framework of EGT for their purposes, but also adopted the biological interpretation of EGT. In particular, they interpret the replication of strategies as a process in which the success of these strategies causes the differential reproduction of strategy bearers, and hence the differential reproduction of strategies themselves.[9]

Yet such a direct import of the biological interpretation is difficult in many domains of the social sciences. These difficulties spring from many sources; I just want to sketch three reasons here. First, while animals largely exist on the subsistence level, humans mainly do not. It is consequently much less clear what the causal effect of, say, adherence to norms is for survival and reproduction in humans, than what the causal effect of daily competition for food, shelter and mating opportunities is for survival and reproduction in non-human animals.[10] Secondly, while it may be plausible that certain basic animal behaviours are encoded in ways that can be inherited through reproduction (albeit of course not in the asexual, breeding-true way assumed in the simplest models), it is much less clear that complex human behavioural characteristics, like compliance with norms, can. Thirdly, the speed of cultural evolution is often much higher than human reproduction. Conventions in small groups, for example, can emerge or change within days, thus making reference to player reproduction inadequate. For these as well as other reasons, strategy replication often has to be thought of in ways independent of player reproduction.

A new interpretation of strategy replication

For the above reasons, some social scientists have accepted that in the social realm, more often than not, strategy replication does not operate through strategy bearer

[9] A notable example is Axelrod (1980), the first paper in a social science journal that uses EGT. He suggests a straightforward adoption of the biological interpretation: 'we simply have to interpret the average payoff received by an individual as proportional to that individual's expected number of [truly-bred] offspring' (Axelrod 1980, 398). A more recent example is Binmore et al. (1995), which interprets payoffs as the players' probability of death.

[10] There are important exceptions to this claim. The selection of social institutions, like firms or markets, seems to fit the biological pattern reasonably well. Interestingly enough, this is largely the playing field of evolutionary economists, while EGT has concentrated on the more complicated cases of evolution of preferences, conventions, and norms.

reproduction. Yet this raises the issue how strategy replication could be realised. Strategy replication does not occur in a vacuum, but requires mechanisms detailing how strategies proliferate. As I will argue now, this poses a problem of its own: however this mechanism is specified, it requires inter-personal comparisons. Without the recourse to the fitness notion, which is bound to the reproduction mechanism, these comparisons may be more problematic.

Examples of such mechanisms are various imitation and experimentation interpretations of EGT models. Under the *imitation* interpretation, players occasionally sample other players in the population, and learn about their strategy and the payoff they realised in the last round. They then switch their strategies according to the following rule: if in a population with state $x(t)$ the agent i's payoff is $u_i(x)$, and the agent samples an agent j with payoff $u_j(x)$, the agent switches with probability

$$q_i = \max\{0, b(u_j(x) - u_i(x))\} \tag{3}$$

(Schlag 1998, 150, cf. also Weibull 1995, 152–161). That is, she retains her strategy if her realised payoffs are greater than that of the sampled player. Otherwise, she adopts the strategy of the sampled player with a probability proportional to the difference between her and the sampled payoff. The function b makes sure that the difference are normalised—i.e. for any payoffs u_i, u_j in the population, $0 \leq b(u_j(x) - u_i(x)) \leq 1$. The imitation interpretation thus makes double use of interpersonal payoff comparisons: once when comparing the payoff of the sampling player with that of the sampled player, and another time when taking into account the payoffs of *all* players (in particular, the maximal and minimal payoffs in the population) when constructing the normalising function b. Note also that the imitation interpretation not only assumes that interpersonal payoff comparisons are possible for the modeller (as the fitness interpretation does), but that it also attributes the capacity to make such interpersonal comparisons (at least between sampler and sampled) to the players: they evaluate others' payoffs as better than their own, and because of this evaluation imitate their more successful opponent.

Under the *experimentation* interpretation, only 'dissatisfied' players change their strategies. 'Individuals with less successful strategies review their strategy at a higher rate than individuals with more successful strategies' (Björnerstedt and Weibull 1996, 162). Player i's review rate $r_{ia}(s)$ when playing pure strategy a against strategy profile s is

$$r_{ia} = \rho_i(u_{ia}(s), s) \tag{4}$$

where ρ_i is an assessment of player i how well she fares playing a against s. The authors do not directly address what player i compares her outcome with, but note that 'ρ_i: $R \times S \to R^+$ is strictly decreasing in its first argument' (ibid.). That is, the higher the payoff of playing a against s, the lower the agent's review rate.

On which judgment can such a function ρ_i be based? The authors stress that it is not presumed that a player 'necessarily *knows* the expected value $u_{ia}(s)$ of her current pure strategy. The only informational assumption is that, on average, the review rate of a-strategists is higher if their expected payoff is lower, *ceteris*

paribus' (ibid., their emphasis). But then the modeller has to construct ρ_i. Given that player i may not even know the payoff for her own strategy, it is unlikely that she will know payoffs for other strategies, unless she observes other players' payoffs. Yet this again requires the comparison of payoffs across different players.

The experimentation interpretation thus makes twofold use of interpersonal payoff comparisons: first when comparing the population's payoffs in ρ_i, and second when constructing the scaling part of ρ_i (in a similar way as the function b in the imitation interpretation) to relate payoffs to the probability of review (which necessarily is interpersonally comparable). Despite the above quotation, not only the modeller, but the player herself must be able to make these comparisons, otherwise the individualised function ρ_i has no structural interpretation.

The imitation and experimentation interpretations of EGT require inter-personal comparisons. This is because they differ from the biological interpretation of EGT only in the way they interpret strategy replication, while both interpretations agree in they way they present strategy selection.

Recall from "The biological interpretation" that the biological interpretation of selection in EGT required the notions of *relative success* of a strategy, as well as *relative strength* of the causal factors operating on the population frequency, comparing success and strength across different strategies held by different players. Similarly with the imitation and experimentation interpretations: strategies are selected from a real social population of strategy bearers. There is no selecting agent who determines which strategies make it into the next generation. In fact, there is no agent selecting strategies on any sort of subjective criterion at all. Rather, the frequency-dependent properties of individual strategies cause their relative rates of replication. The mechanisms of this causal relation are represented in the imitation rule and the experimentation review rate, respectively. The strategy's relevant properties, and their relative causal strength, are summarised in the strategy's payoffs. Because these causes all operate on the same population property, the payoffs that represent them must be comparable across all individuals of that population.

More generally, models like imitation and experimentation are examples of modelling selection in material play. Such selection interpretations of EGT model the effect of material play—of what players do, and which outcomes actually obtain—on relative strategy frequencies of a population. That is, they model effects of individual material play on an aggregate property. Yet the causes they model are represented as the individual players' payoffs. Because causes affecting the same property must be comparable in their force, the players' payoffs that represent these causes must be comparable. Therefore, actual play interpretations of EGT models require interpersonal payoff comparisons.[11]

To summarise, social scientists have adopted the biologists' EGT models for their own purposes. Yet the biologists' interpretation of strategy replication often

[11] EGT under the actual play interpretation has also been employed to model *individual* learning. Drawing on the Bush-Mosteller reinforcement learning concept, Börgers and Sarin (1997), for example, modelled an individual agent's learning as an adjustment of the weights of her mix strategy in proportion to her payoffs from past play. Such individual learning models of course do not require interpersonal comparisons, as they model an individual, not a social process.

does not suit social scientists' goals. Thus they replaced it with a number of alternative mechanisms that operated through imitation, experimentation or similar. All of these mechanisms require that the theory compare payoffs across players. Thus, evolutionary game theory under the material play interpretations requires interpersonal payoff comparisons.

Digression: are inter-personal payoff comparisons problematic?

The above conclusion contrasts evolutionary game theory with classical game theory, as used in economics and the social sciences. Classical game theory only requires that players compare their own payoffs across different strategies, not that players compare different players' payoffs. This is commonly seen as a strength of classical game theory. It is therefore noteworthy that EGT requires interpersonal payoff comparison under the material play interpretation, as it contradicts the common perception that evolutionary game theory makes weaker epistemic demands. In terms of interpersonal comparisons, alas, evolutionary game theory under this interpretation makes *stronger* assumptions than classical game theory does.

Nevertheless, interpersonal payoff comparisons need not necessarily be a problem for applying evolutionary game theory in the social sciences. First, there may be social situations where the biological interpretation of strategy replication applies, and hence where strategies replicate through player reproduction. Selection of social or commercial institutions, as mentioned in footnote 10, may fall under this category. In that case, the payoffs can be interpreted as individual fitness, which is inherently interpersonally comparable. Second, there may be social situations, even under a non-biological interpretation of strategy replication, where legitimate interpersonal payoffs comparisons can be made. Most people, after all, have some intuition about how well off they are vis-à-vis others. This is particularly true in situations where people compete for a single, enumerable resource, like for example money.[12] Thus, it can be argued that at least in these situations, even the imitation and experimentation models can be used.

Yet these considerations also show the limits of interpersonal payoff comparisons. In many situations, the consequences of interactions cannot be adequately represented as the measure of a uniform objective quantity. Instead, consequences often consist of bundles of objective consequences (e.g. money/work-hour pairs), or consequences include psychological qualities (risk, damage to others). This is particularly true for models concerning the evolution of preferences, conventions, and norms—areas that have been the main focus of EGT applications in the social sciences.

It should be noted here that the unifying term 'payoff' masks crucial conceptual differences, which in turn affect what exactly is interpersonally compared. In the standard utility case, payoffs capture subjective evaluations of strategy outcomes. A

[12] See for example Alexander (2007), who calls them 'public goods', because they are publicly comparable.

number of different formalizations of such measures exist, but they all have in common that they are not *inter*personally comparable. To compare utility numbers of one individual with another, under any of these measures, simply does not have meaning at all. But how to construct an appropriate measure? The assumptions necessary for modelling such comparisons are a stone of continuing contention. It seems very difficult to make generalisable claims about when agents are capable of making interpersonal comparisons, and hence when modellers are justified making assumptions of this sort (for an overview, see Hammond and Fleurbaey 2004). Under the biological interpretation of EGT, payoffs capture a strategy's properties that differentially contribute to the individual's reproduction. As argued above, such a fitness interpretation seems often legitimate in biology, and sometimes also legitimate in the social sciences. In the non-biological material selection interpretations, payoffs capture a property of the strategy that contributes to the tendency of individuals to imitate or to review it. Such an interpretation is more problematic than the biological one, as these causal factors do not operate through the reproduction of the individual anymore. Rather, they operate through the uniform assessment of all players in the population. Thus, the payoffs do not represent anything relevant for the individual player anymore, but rather a general 'attractiveness' or 'replicative power' of a strategy. Whether such an interpretation can be justified may depend on individual cases. But generally speaking, it seems a rather non-promising case, not least because a model that explains replication with 'replicative power' threatens to be trivial.

To be sure, this is not to claim that devising a measure that allows interpersonal comparisons is impossible. I am simply claiming that there currently exist no agreed upon way how to measure and interpret it. So the conclusion is that social scientists employing EGT with a material play interpretation will have to be mindful of the appropriateness of the payoff comparisons for the cases studied. A simple transfer of the biological interpretation, which so often is found in the social science literature, will not do.

A new interpretation of strategy selection

For those cases where interpersonal comparisons pose a problem, a solution may be to re-interpret EGT in general, and strategy selection in particular, as a theory of individual mental processes. Under this interpretation, all references to payoffs of others in a given environment are understood counterfactually as the payoffs that one would get in that environment, if one adopted the other's strategy. For example, if a player knows the payoffs of each strategy profile, and knows the frequency with which strategies are played in the population, she can compare the expected payoffs of these strategies based solely on her own preferences. Having compared the strategies according to her own preferences, she can then choose that strategy that is either better than the current strategies, or a best reply to her belief about the frequencies in the population. Variants of such models have been proposed by Sugden (1986), in Kandori's et al. (1993) 'stochastic fictitious play', and in Young's (1993) 'adaptive play'.

Take for example Young's (1993) model. He defines play at time t as the strategy-tuple $s(t) = (s_i(t),\ldots, s_n(t))$, consisting of each player's strategy choice at time t. At period t, each player samples the past play of a certain number of past periods. From this sample, the player constructs strategy tuple s_h by weighing the past play in some way. Strategy tuple s_h constitutes her estimate how other players will play in the next period. Thus, for the next period, agent i chooses s_i as the best reply to s_h.

Young's model is an example of what I call a *mental play* interpretation of EGT. What is relevant for a certain strategy to be selected no longer is the effect of actual interaction in a real population, but rather the consequence of an individual player evaluating various options, based on her subjective value criteria and her beliefs what her opponents will play. She forms these beliefs from her perception of and through reasoning about others' past play. She chooses by mentally representing her various options in the anticipated environment, figuring out the consequences of these counterfactual scenarios and choosing the one with the outcomes she values better or best.

This distinguishes mental play models from material play models. In the latter, material interaction in a genuine population causes differential strategy replication. That is, A interacts with B, observes B's payoff, compares it to his own, and is caused to adopt a certain strategy through some imitation of experimentation process. In the former, actual interaction influences players' beliefs about strategies, payoffs, population frequencies, etc. Based on these beliefs, players later pick a strategy from those that they have a mental representation of. That is, A interacts with B, but only observes his own payoff. He then imagines playing different strategies in the same context, comparing the actual and imagined outcomes according to his own judgment, and selecting the best one. He thus *anticipates* and chooses strategies accordingly. Under the mental play interpretation, EGT only models the effect of past interaction on strategy selection mediated by deliberation and mental representation.

Consequently, because the causal relation is between interaction and individuals' mental attitudes, no interpersonal payoff comparison is necessary. Players only observe their own payoffs from past play, and this affects only their own attitudes towards future play. Effects on aggregate properties are not directly modelled.

There are a number of problems with this interpretation. First, it re-introduces relatively strong knowledge and cognitive capacity assumptions back into evolutionary game theory. In particular, the player is assumed (1) to be able to discern which strategies players play, (2) to know the payoffs for all strategy profiles present in the population, and able to evaluate them; (3) to know the frequency with which strategies are played in the population; and (4) to be able to compute expected utilities from this knowledge. One motivation for social scientists to introduce evolutionary game theory was exactly to avoid strong assumptions of this kind. Thus, arguably, part of the desirability of evolutionary game theory is undone through the mental play interpretation.

Second, the mental play interpretation weakens the rationale for the replicator dynamic and similar monotone dynamics.[13] In these dynamics, any strategy with an

[13] Payoff-based dynamics are called *monotone* if and only if the difference between individual and average payoff has a monotone increasing effect on that strategy's relative growth rate.

above average (or above median) payoff has a positive growth rate. Yet in the mental play interpretation, it is implausible that agents will play strategies that are worse than the best strategy. Instead, the interpretation naturally implies that agents will choose a best reply, eliminating all but the highest-achieving strategy. The resulting dynamic is thus considerably different from monotone dynamics. Given that many authors consider evolutionary game models to be characterised by monotone dynamics (e.g. Fudenberg and Levine 1998, 51–53), one may wonder whether the mental play interpretation should be considered as a part of evolutionary game theory at all.

Third, as Kuhn (2004, 12) argues, if one assumes that players know the strategies present in the population (as well as their payoffs), why not assume that they know all possible strategies and their payoffs, and choose a best reply to them? This would of course lead back full circle to standard game theory and its solution concepts, eliminating the particular approach of evolutionary game theory altogether. One way to justify the restriction to strategies present in the population is to maintain that players do not know strategy payoffs, but rather infer this from observing payoffs to others in the population. Yet as Kuhn correctly points out, this argument would again require that players are capable of comparing others' payoffs to their own, and hence assume interpersonal payoff comparability. The mental play account's insistence on considering only strategies present in the population thus does not have a stable justification; in combination with the lost rationale for monotone dynamics, the question arises why the mental play interpretation should at all be considered part of evolutionary game theory. Rather, upon reflection, it seems much closer to classical game theory, or a bounded rationality version thereof.

These problems are individually relevant, and together form a reason to be sceptical about the success of the mental play interpretation of EGT. Yet I want to approach this problem from a different angle. In the following, I argue that evolutionary game theory, when interpreted as mental selection, falls outside of the class of theories that model natural selection.

EGT and natural selection

Social scientists who adopted EGT often emphasise its ability to model a particular kind of selection. For example, Samuelson says that

> Economic theory is now routinely described as assuming … that some *process of selection*…will cause us to observe people who act as if they are maximizing. … Evolutionary game theory brings the evolutionary portion of these arguments out of the background. (Samuelson 1997, 51, emphasis added)

And Mailath argues that game theoretic

> evolutionary dynamics do not build in any assumptions on behavior or knowledge, other than the basic *principle of differential selection* – apparently

> successful behavior increases its representation in the population, while unsuccessful behavior does not. (Mailath 1998, 1355, emphasis added)

It remains somewhat unclear from these quotes what notion of selection they have in mind. However, they clearly contrast it with individual maximising choice: while it looks 'as if' people were maximising, really other forces are at play. This becomes even clearer in a quote from Binmore:

> Maynard Smith's book *Evolution and the Theory of Games* directed game theorists' attention away from their increasingly elaborate definitions of rationality. After all, insects can hardly be said to think at all, and so rationality cannot be so crucial if game theory somehow manages to predict their behavior under appropriate conditions (Binmore, foreword in Weibull 1995, x).

Insects do not have cognitive abilities to perform a maximising procedure. The question is what these non-mental forces are that Binmore seeks to employ for the benefit of the social sciences.

Recall from "Transfer into the social sciences" that economists often employed EGT in a recovery effort—to justify solution concepts, in particular the Nash equilibrium, and to offer plausible equilibrium selection and refinement criteria. How well EGT serves these purposes depends less on its formal properties, but rather on the way it is interpreted. It seems that many believe that the foundations of classical game theory are redeemed by the convergence results from EGT (Sugden 2001), popularly understood to be saying that the relevant Nash equilibria emerge as the survivors of natural selection processes. Understood as modelling natural selection, EGT taps into the explanatory successes of the concept in biology. Through this unificationist move, theorists hope to justify the Nash equilibrium as result of natural selection. Natural selection operates as a well-supported mechanism that helps prop up some of economics' own, more difficult principles.

While it may not apply to all social scientists, it appears therefore that at least many of them seek to interpret the social phenomena they model with EGT as instances of natural selection. Hence, the question whether EGT can be interpreted as a theory of natural selection is of significance for social scientists.

Various accounts have been offered on what constitutes a theory of natural selection, (also called a selection-type theory). At the core of these summaries are (1) the variation of individuals' traits, (2) the causal relationship between these traits and the individuals' differential rates of reproduction, and (3) the inheritability of these traits (for overview and discussion of these accounts, see Godfrey-Smith 2007). In this section, I will focus on the second of these necessary conditions. Does EGT, under its various interpretations, offer a causal relationship between the interaction of traits (i.e. strategies) and the differential replication of these traits? Answers to these questions will help to clarify which of these interpretations are capable of modelling natural selection.

Under the biological interpretation, the answer is—unsurprisingly—yes. In the typical biological cases, the causal link between interaction and differential replication of strategies is suggested to go through the differential reproduction of

the strategy bearers. Outcomes of interactions affect individual players, and through this affect individual reproductive fitness. Player reproduction then realises strategy replication. Of course, questions concerning the stable effect of payoffs on fitness and concerning the relation between reproduction and trait replication remain. But despite these questions, the model offers a causal process from the interaction of traits to the differential replication of these traits, and hence satisfies this necessary condition for modelling natural selection.

Under the material play interpretation the answer is also yes. In imitation players copy the strategies of those who did comparatively better in past play. In experimentation, players abandon a strategy if they did worse than others. In these cases, it is the effects of actually playing a certain strategy that causes the proportional increase of that strategy in the population. But now the causal link between interaction and replication is established through players adopting, abandoning, or adjusting their strategy as an effect of their previous interaction. The ways interaction and replication is linked are different, but the causal process between the two is explicitly modelled.

The same cannot be said about EGT under the mental play interpretation. Under that interpretation, what is relevant for a certain strategy to be adopted is no longer actually interaction in the real population, but that individual players counterfactually consider how well their strategy options would serve them. The theory focuses on players' individual anticipations, not on their material interactions. Actual interactions only have an influence on some of these anticipations. The causal link between interaction and strategy replication is thus weakened, and mediated by several potentially disrupting instances. For example, choice of different reasoning principles or heuristics may lead to different beliefs about strategies, strategy outcomes, etc., even when based on the same actual interactions. Similarly, choice of different decision principles or heuristics will lead to different influences of the same beliefs on intentions and choices. Further, while under the material play interpretation an interpersonally comparable evaluation secured the causal power of certain interactions on strategy replications, the choices of individual players now depend on subjective evaluations. Different players may thus have different reactions to identical interactions. I therefore conclude that there is no actual interaction, and hence no causal forces emanating from it, which directly influences differential replication of strategies. The mental selection version of evolutionary game theory cannot anymore model the causal process that links interaction and replication.

EGT under the counterfactual interpretation therefore does not satisfy one of the necessary conditions of selection-type theories. It falls back into the category of a theory of reasoning equilibria, and thus is closer to standard versions of game theory than to the biological interpretation of EGT.

Conclusion

The recent transfer of EGT into the social sciences poses an interesting challenge: how and to what extent does this transfer require a re-interpretation of the formal

framework. I have distinguished three basic kinds of responses: First, acceptance of the main tenets of the biologists' interpretation; second, replacement only of the biologists' interpretation of strategy replication; third, replacement of both the biologists' interpretation of strategy replication and of strategy selection.

Each of these responses faces its own problems, which limit the application of the EGT framework in some way. The full acceptance response is hampered by the fact that strategy replication often cannot be sensibly interpreted as strategy bearer reproduction in this domain. The strategy replication re-interpretation stays within the category of material play models, but all of these models require the interpersonal comparability of strategy payoffs. Because meaningful interpretation to such comparisons is not an easy task for many situations in the social sciences, the applicability of material play models is also limited. Finally, the reinterpretation of both replication and selection leads to mental play models. These solve the comparability issue, but are incompatible with natural selection interpretations of EGT. To the extent that social scientists seek such a natural selection interpretation, they thus face a dilemma: either face the challenge that interpersonal comparisons pose, or give up on the natural selection interpretation. For some situations, one horn of this dilemma may be more acceptable than the other, but both often remain disadvantageous. Between these two horns, the use of EGT to the social sciences is significantly diminished.

Of course, the conceptual tensions between the formal framework of EGT, its employment in the social sciences, and the three interpretations do not in principle rule out the application of EGT models in the social domain. Rather, my analysis shows potential problems of applicability, and pleas both for greater awareness of the specific purposes of EGT modelling in the social sciences, as well as for greater sensitivity to the underlying microstructure on which the evolutionary dynamics and other EGT solution concepts supervene.

Acknowledgments Earlier versions of this paper were presented at LOFT'08 in Amsterdam on July 3rd, 2008 and at the workshop Evolution & and the Human Sciences, Helsinki, on November 13th, 2009. I thank the participants of these sessions for their comments. Particular thanks are due to Aki Lehtinen for very insightful discussions and to two anonymous referees for helpful comments.

References

Alexander JM (2007) The structural evolution of morality. Cambridge University Press, Cambridge

Axelrod R (1980) More effective choice in the Prisoner's Dilemma. J Conflict Resolut 24:379–403

Binmore KG, Samuelson L, Vaughan R (1995) Musical chairs: modelling noisy evolution. Games Econ Behav 11:1–35

Björnerstedt J, Weibull JW (1996) Nash equilibrium and evolution by imitation. In: Arrow KJ, Colombatto E, Perlman M, Schmidt C (eds) The rational foundations of economic behavior. St. Martin's Press, New York, pp 155–181

Börgers T, Sarin R (1997) Learning through reinforcement and replicator dynamics. J Econ Theory 77:1–14

Fudenberg D, Levine DK (1998) Theory of learning in games. MIT Press, Cambridge

Gintis H (2000) Game theory evolving. Princeton University Press, Princeton

Godfrey-Smith P (2007) Conditions for evolution by natural selection. J Philos 104:489–516

Grüne-Yanoff T (2011) Models as products of interdisciplinary exchange: evidence from evolutionary game theory. Stud Hist Philos Sci 25(2):1–19

Hammond P, Fleurbaey M (2004) Interpersonally comparable utility. In: Hammond P, Barbera S, Seidl C (eds) Handbook of utility theory, Vol. 2: extensions. Kluwer Academic Publishers, Dordrecht, pp 1181–1285

Kandori M, Mailath G, Rob R (1993) Learning, mutation, and long run equilibria in games. Econometrica 61:29–56

Kuhn ST (2004) Reflections on ethics and game theory. Synthese 141:1–44

Mailath GJ (1998) Do people play nash equilibrium? Lessons from evolutionary game theory. J Econ Lit 36(3):1347–1374

Maynard Smith J (1982) Evolution and the theory of games. Cambridge University Press, Cambridge

Millstein RL (2006) Natural selection as a population-level causal process. Br J Philos Sci 57:627–653

Rosenberg A, Bouchard F (2010) Fitness. In: Zalta EN (ed) The Stanford encyclopedia of philosophy, Fall 2010 Edition. URL http://plato.stanford.edu/archives/fall2010/entries/fitness/

Samuelson L (1997) Evolution and game theory. J Econ Perspect 16(2):47–66

Schlag K (1998) Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. J Econ Theory 78(1):130–156

Sugden R (1986) The evolution of rights, cooperation, and welfare. Basil Blackwell, Oxford

Sugden R (2001) The evolutionary turn in game theory. J Econ Methodol 8:113–130

Weibull J (1995) Evolutionary game theory. MIT Press, Cambridge

Wimsatt WC (1980) Reductionistic research strategies and their biases in the units of selection controversy. In: Nickles T (ed) Scientific discovery: case studies. D. Reidel, Dordrecht, pp 159–213

Young P (1993) The evolution of conventions. Econometrica 61:57–84