The evolutionary puzzle of human cooperation is a red herring

Abstract

An influential line of theoretical modeling and analysis is directly motivated by describing human cooperation as an "evolutionary puzzle." This puzzle paradigm asks how human cooperative behaviors can be understood as the outcome of an evolutionary process. We reveal a critical flaw in the puzzle paradigm for the study of human cooperation: This flaw stems primarily from the fact that puzzle models define cooperation in terms of fitness costs and benefits, which starkly contrasts with the standard evolutionary approach of deriving the fitness costs and benefits of a behavior from a consideration of how the behavior interacts, via life-history, with evolutionary selection. As a consequence, puzzle models exist in a closed theoretical loop, rendering them incapable of making meaningful predictions or providing any insight into the origin or existent patterns of human cooperation. We also discuss why the puzzle paradigm cannot be salvaged by conceiving of the evolutionary process as cultural rather than genetic. We conclude that this approach to human cooperation is a red herring but that valuable ideas for future work can be gained from questioning its tacit assumptions.

Keywords. altruism, cultural evolution, evolutionary game theory, genetic evolution, human cooperation, phenotypic gambit

1. Introduction

Our aim of this paper is to offer a critique of a popular research paradigm. We shall refer to it as the *puzzle paradigm* as its use tends to be motivated by an assertion that human cooperation is an evolutionary puzzle. We use four influential papers as examples of the puzzle paradigm. As described in detail in section 2, these papers are characterized by a two-step sequence of mathematical modelling: (1) Human cooperative behavior is modeled as an altruistic strategy in a giving game. The puzzle is that although humans seem to cooperate a great deal, altruistic strategies in a giving game should be unsuccessful in any evolutionary process whereby replication is based on game payoffs. (2) The giving game is then embedded in an iterated social context. This embedding is claimed to resolve the puzzle of human cooperation through a demonstration that successful strategies may involve altruistic moves.

We shall argue that the puzzle paradigm is implicitly employing a research strategy from evolutionary biology called the phenotypic gambit. In section 3 we describe the phenotypic gambit and discuss how it is not justified unless researchers take care to satisfy several criteria for its use. Two examples are given of such careful use of the phenotypic gambit in evolutionary biology. In section 4 we show how puzzle paradigm research compares unfavorably with these examples and discuss the problems that ensue. In section 5 we question the premise that human cooperation should be studied as an evolutionary puzzle. We discuss how questioning this premise opens up more promising avenues for research.

2. The puzzle paradigm

We have selected four influential papers (Boyd et al. 2003; Nowak & Sigmund 2005; Hauert et al. 2007; Boyd et al. 2010) as canonical representatives of the puzzle paradigm. These papers, which appeared in the highest ranked general science journals, study how cooperation in humans may have evolved in tandem with costly punishment mediated

through a variety of mechanisms. Each paper begins by asserting or implying that human cooperation is an evolutionary puzzle:

> "Unlike any other species, humans cooperate with non-kin in large groups. This behavior is puzzling from an evolutionary perspective […]" (Boyd et al. 2003).

> "[…] human societies, are organized around altruistic, cooperative interactions. How can natural selection promote unselfish behaviour?" (Nowak & Sigmund 2005).

> "All human populations seem willing to use costly punishment to varying degrees, and their willingness to punish correlates with the propensity for altruistic contributions. This raises an evolutionary problem […]" (Hauert et al. 2007).

> "Humans are a uniquely cooperative species […] Recent theoretical studies provide an evolutionary explanation for such cooperative behavior […] There are two important problems with this explanation of human cooperation" (Boyd et al. 2010).

These quotes highlight the two distinct but related research questions the puzzle paradigm seeks to answer: Why are humans unique among species with regard to cooperation, and how can cooperative behaviors be selected for by an evolutionary process? The second question is precisely the central question of the branch of theoretical biology that studies the evolution of altruism (e.g., Fletcher & Doebeli 2006).

2.1 The first step: Modeling cooperation as altruism in a giving game

In the puzzle paradigm cooperative behavior is taken as a synonym of altruism and defined as a behavior where an individual pays a personal fitness cost to bestow a fitness benefit upon others. We will refer to a situation that offers the possibility for such an altruistic act as a "giving game". Thus, in the puzzle paradigm cooperative behavior is

modeled as an altruistic act in some version of a giving game as indicated by the following quotes.

> "There are two behavioral types, contributors and defectors. Contributors incur a cost *c* to produce a total benefit *b* that is shared equally among group members. Defectors incur no costs and produce no benefits." (Boyd et al. 2003)

> "In the simplest model, the altruistic act consists in conferring a benefit *b* on the recipient at a cost *c* to the donor." (Nowak & Sigmund 2005)

> "Those who participate can decide whether or not to contribute an investment at a cost *c* to themselves. All individual contributions are added up and multiplied with a factor $r > 1$. This amount is then divided equally among all participants of the public goods game." (Hauert et al. 2007)

> "Cooperation costs the cooperator *c* and benefits each member of the group *b/n* (*b*>c> *b/n*)." (Boyd et al. 2010)

From these quotes it is clear that cooperation is defined in terms of the payoffs in this giving game. Moreover, these payoffs are assumed to be in the currency of fitness. In other words, the strategy that incurs the largest net benefit will increase most in frequency. A "cooperative" strategy of paying a cost to give a benefit to others should therefore be outcompeted by a "non-cooperative" strategy. The emergence and persistence of high levels of cooperation under these conditions should not be possible.

2.2 The second step: Embedding the giving game to resolve the puzzle

The puzzle paradigm's research goal is to resolve the proposed puzzle by showing that high levels of cooperation can emerge and persist under the additional assumption that the giving game is embedded in an iterated social context, in other words a meta-game. This

embedding varies from paper to paper and is rather lengthy to describe. For instance, the embedding of Boyd et al. (2003) amounts to the following four steps: (1) structuring the population into groups, with a certain amount of migration between groups; (2) adding a third strategy type that both cooperates *and* punishes non-cooperators in their group, at certain costs to themselves and the punished parties; (3) letting relative replication of strategies depend on the total payoffs obtained from the giving game and punishments; (4) letting groups meet each other in pairwise competition, with the probability of winning decided by the frequency of cooperation in each group, and with the winning group replacing the losers. Given this embedding it is then demonstrated that, as meta-game strategies replicate, cooperative behavior in the giving game and punishing behavior in the punishment game are both sustained at high levels.

The other three puzzle paradigm papers in our sample similarly embed the simple giving game in a complex meta-game, each with some particular features: Nowak and Sigmund (2005) deal with reputation and indirect reciprocity, Hauert et al. (2007) with optional participation, and Boyd et al. (2010) with coordination and economies of scale.

3. The phenotypic gambit

A "gambit" is a chess opening in which a piece is sacrificed in order to gain what is hopefully a better position. The phenotypic gambit was perhaps first made by Fisher (1930) in an attempt to understand the evolution of sex ratios. The term phenotypic gambit was coined in Grafen (1984). We quote part of his definition:

> "The phenotypic gambit is to examine the evolutionary basis of a character as if the simplest possible genetic system controlled it: as if there were a haploid locus at which each distinct strategy was represented by a distinct allele, as if the payoff rule gave the number of offspring for each allele, and as if enough mutation occurred to allow each strategy to invade." (Grafen 1984)

This quote emphasizes the simplification of genetic inheritance mechanisms, while perhaps deemphasizing the simplification of the complex and context dependent mapping from genotype to fitness via phenotype. We encourage readers to read Grafen's introduction of the term in its entirety, as well as other related accounts of when and why such a gambit is justified (Smith 1978; Dawkins 1982 pp. 30-54). Here we will do our best to paraphrase these accounts.

3.1 The sacrifice of complexity

The basic idea of the phenotypic gambit can be expressed as follows: Instead of starting with a genetic replicator and asking whether its frequency will increase or decrease in a population as a result of evolution, the researcher starts with a phenotype and asks the same question. In order for the frequency of this phenotype to be affected by evolution at all, the variation in the phenotype must be to some extent heritable (i.e., co-vary with underlying alleles at least at one, but possibly many, genetic loci). The details of this genetic heritability may be complicated by various factors such as pleiotropy, epistasis, and linkage disequilibrium. The first thing the phenotypic gambit puts on the chopping block is this genetic complexity. Replacing it is a set of haploid alleles at a single locus, which correspond one-to-one with the set of possible phenotypes or strategies.

The question of which phenotypes are actually possible is dependent on the genetic mechanisms by which new phenotypes arise. In general, this phenotypic variation will be constrained by the phylogenetic history of the organism, the ontological process by which the phenotype develops, and the bio-physical processes by which the phenotype manifests. The phenotypic gambit also largely ignores these constraints by leaving the choice of strategies to be considered at the discretion of the researcher. Choosing the set of strategies to be considered is a delicate task, fraught with the potential for implicit assumptions. In addition,

not only the set of available strategies but also the manner in which new strategic variants (mutants) derive from pre-existing types radically impacts evolutionary outcomes (e.g. van Veelen et al. 2012).

To summarize, the phenotypic gambit sacrifices several layers of complexity at a considerable risk of over-simplification. The gain comes in the form of access to powerful theoretical tools, to be described next.

3.2 The analysis enabled by the sacrifice of complexity

The notion of evolutionarily stable strategies (ESS; Smith & Price 1973) provides a clear and intuitive way of thinking about evolutionary outcomes and making evolutionary predictions. Taylor and Jonker (1978) formally linked the ESS to the equilibria of genetic dynamical systems of the kind described in the above quote from Grafen (1984). In other words, the particular assumptions of the phenotypic gambit empower a theorist to forgo an explicit investigation of genetic dynamics, which may be intractable, and focus instead on a relatively simple ESS analysis. However, even after this simplification there remains the task of deriving an appropriate fitness function.

A fitness function takes as inputs the frequencies of each strategy in the current population and gives as output the relative rates of growth or decline for each strategy. The derivation of the fitness function constitutes the core of a phenotypic gambit analysis. If the strategies being considered have been empirically observed, and the fitness of each strategy is known, this function may be fit to data. Typically though, the phenotypic gambit is used to provide an evolutionary explanation for the existent variation, both inter and intra species, in light of the strategic situation created by the interaction of various ecological factors. In such cases empirical fitting of the fitness function may be infeasible. In some cases the behaviors considered are intrinsically linked to individual fitness. The derivation of a fitness function

from the given strategy set and life history assumptions is then unambiguous, though possibly mathematically complex. In more challenging cases, strategies have ambiguous impacts on fitness and theoreticians are left with the hard choice of either making further simplifying assumptions, until the derivation of fitness becomes unambiguous (e.g., Higginson et al. 2012), or attempting to work with a general broad class of possible fitness functions (e.g., Pen & Taylor 2005).

3.3 The caveats of the phenotypic gambit

Once derived, the fitness function manifests the essence of an interesting trade-off or strategic social interaction. However, even if the fitness function is ecologically valid there is no guarantee that the implications of this fitness function under the assumptions of the phenotypic gambit will still hold when these assumptions are relaxed. A researcher employing the gambit hopes, but does not know, that the dynamics and equilibria resulting from the simplified strategy set and genetic system are a good indication of the dynamics and equilibria of the actual phenotype, which is determined by a complex genetic system interacting with ontogeny via messy bio-mechanical processes and evolving under significant phylogenetic constraints.

Some experts have expressed some doubt of the usefulness of ESS analyses when the actual genetic complexity is greater than the model allows for (e.g., Taylor 1979). There have also been more passionate critics of the phenotypic gambit (Lewontin 1979; Gould & Lewontin 1979; Gould 1982), as represented by the following quote:

> "An adaptationist programme… based on faith in the power of natural selection… proceeds by breaking an organism into unitary 'traits' and proposing an adaptive story for each considered separately…We criticize this approach and attempt to reassert a competing notion that organisms must be analysed as integrated wholes." (Gould & Lewontin 1979)

When the dust settled after this debate the mainstream conclusion was not that organisms must be analyzed as integrated wholes but at least that care needs to be taken when identifying the particular trait that is to be the focus of an adaptationist analysis (Dawkins 1982 pp. 33-54; Grafen 1984), though see Andrews et al. (2002). Next we shall present two careful applications of the phenotypic gambit.

3.4 The phenotypic gambit in practice

To provide clear examples to contrast with puzzle paradigm models we discuss two specific applications of the phenotypic gambit. One of these (Hamilton 1967) was written prior to the formalization and entrenchment of the phenotypic gambit and is thus very explicit about the various assumptions made. The second one (Kokko & Jennion 2008) was written recently within a research culture where many of the critical assumptions of the phenotypic gambit have become largely tacit. We shall now discuss how, despite the decades separating these papers, they both share the same set of critical features ensuring that the gambit is not made in vain: The phenotype is well-defined and can reasonably be studied in isolation; a fitness function is derived from interaction between the phenotype and relevant ecological factors; and the theoretical results yield interesting predictions that can be tested in empirical work.

3.4.1 The phenotype is well-defined and can reasonably be studied in isolation

The first critical feature of a successful phenotypic gambit is a clear definition of the phenotype to be explained. For Hamilton (1967) the phenotype of interest is the sex ratio at birth, an unambiguous concept. For Kokko and Jennion (2008) the focal phenotype is parental investment, specifically operationalized as the ratio of time spent caring for existent offspring versus time spent seeking opportunities to create more offspring, again unambiguous.

The phenotypes considered by Hamilton (1967) and Kokko and Jennion (2008) seem reasonable to study in isolation. The "decision" of which sex ratio to have is made by every (female or male depending) organism in a species every time it reproduces and this has direct and relatively consistent fitness consequences at all times and places. The same is true of the "decision" between investing in current existent offspring and investing in making more offspring (although environmental changes can have significant impact here and so the theory applies only when the relevant aspects of the environment are sufficiently stable). Barring some deeply engrained phylogenetic or ontological constraints on variation, the fact that these traits have direct and consistent fitness effects for every individual in every generation suggests that the phenotypic gambit will produce reliable results.

3.4.2 A fitness function is derived from interaction between the phenotype and relevant ecological factors

The next critical feature of a successful phenotypic gambit is a clear hypothesis about what ecological factors should be considered, and a derivation of a fitness function parameterized by these ecological factors. For instance, Hamilton (1967) hypothesized that sex ratios would be affected by localized mate competition among males coupled with population wide competition among females for breeding sites. The hypothesis of Kokko and Jennion (2008) is that sexual divergence in parental care might be explained by differential mortality rates when providing parental care versus searching or competing for mating opportunities, by differences in parental certainty, and by feedbacks between these two factors mediated through operational sex ratio. The researchers of these behaviors derived fitness functions by making careful arguments about how the relevant ecological factors should interact with strategies to affect the number of surviving offspring under different life histories. Although the specific life history model of Kokko and Jennion (2008) is largely

implicit, those familiar with the field can reconstruct the explicit model from which the various fitness functions are derived. For example, they implicitly assume that male and female parental care can evolve independently so that males and females effectively form independent populations.

In both papers the primary goal is to explain variation across species, though within species variation is briefly discussed in both. Kokko and Jennion are careful to point out that the factors that explain individual variation within species need not be the same factors that explain between-species variation.

3.4.3 The theoretical results yield interesting predictions that can be tested in empirical work

When a fitness function has been derived, researchers can determine evolutionarily stable strategies (although in 1967 when the paper of Hamilton was written the term ESS had not yet been fully developed; foreshadowing future developments he phrased his arguments in terms of "unbeatable" sex ratios). Under the crucial assumption that the evolutionary process is close to equilibrium, this yields some predictions about the world. The model of Hamilton predicts that the sex ratio of a species will be female biased if females compete globally within a population for breeding sites but males compete locally (and hence often with their siblings) for mating opportunities, and that the degree of this bias will increase with the degree to which male offspring compete and breed with their siblings. These qualitative predictions were compared with empirically observed patterns, allowing Hamilton (1967) to conclude that "in its direction, the effect that Wylie and Jackson have independently reported accords with the theory." Similarly, Kokko and Jennion (2008) drew connections between previous empirical observations and their model predictions of parental investment under different combinations of the relevant ecological factors. In addition, the authors suggested specific future empirical work based on the models presented.

It is a very valuable feature of a theory that it produces predictions (qualitative or quantitative) that can be empirically refuted or verified. Theories with this feature have the possibility of becoming part of a larger body of interlocking theoretical, observational, and experimental work, as has been the case for both of these examples.

4. Comparing the puzzle paradigm to the phenotypic gambit

Thus far we have described the puzzle paradigm approach to studying human cooperation (section 2) and the research strategy in evolutionary biology known as the phenotypic gambit (section 3). We shall now compare the two.

Recall that the phenotypic gambit is to examine the evolutionary basis of a character as if the simplest possible genetic system controlled it (Grafen 1984). This is exactly what the puzzle paradigm does: It studies the evolution of cooperative vs. non-cooperative behavior as if these behaviors were controlled by distinct alleles at a haploid locus (although the "as if" part is typically not stated so explicitly). Because of this similarity, it is worth examining how the puzzle paradigm fares with respect to the criteria for successful use of the phenotypic gambit that were outlined in section 3.4.

4.1 The phenotype of the puzzle paradigm is not well-defined and cannot reasonably be studied in isolation

Cooperation is the phenotype of interest in the puzzle paradigm, and is clearly defined within this paradigm as paying a cost to bestow a benefit to others. It is crucial to note that this definition is in terms of the payoff produced by the behavior. Thus the phenotype is not a concrete behavior but rather an abstract category of behaviors organized according to their fitness effects. This leads to two problems: The first problem is that the definition of cooperation in terms of payoffs is ambiguous. The second problem is that strategies defined by payoffs cannot reasonably be studied in isolation.

4.1.1 The definition of cooperation is ambiguous

The puzzle paradigm sets out to study the actual phenomenon of human cooperation. However, it then provides a definition of cooperation that is identical to the definition of altruism used by theoretical biologists, provided that the costs and benefits of the giving game are in terms of individual reproductive success. Within the altruism research paradigm, researchers begin by positing the existence of genes with fitness effects equivalent to cooperation and defection in the giving game, and then attempt to model the evolutionary dynamics of these genes within a given reproductive system, without regard for what the altruistic behavior might be that causes these fitness effect or the causal relationship between gene and behavior (e.g., Hamilton 1963). Without doing the hard work of linking such abstract definitions to any actual behavior, we do not know what actual behavior the models apply to. This is fine if the aim of the research is confined to the purely theoretical domain, but not if the aim is to understand an actual phenomenon such as human cooperation.

In the four papers we have chosen to represent the puzzle paradigm no concrete behaviors are referred to. Behaviors are only described in abstract terms such as “unselfish”, “altruistic”, and “costly”, which are defined in reference to assumed payoffs. However, other papers by some of the same authors have mentioned behaviors that they consider to be cooperative, such as giving blood (Boyd & Richerson 1998) and giving to charity (Bowles & Gintis 2002). The hard work then remains to show that these concrete behaviors satisfy the abstract definition (e.g., that giving to charity reduces one’s fitness), and that they can reasonably be studied in isolation.

This problem is exacerbated when puzzle paradigm research (unlike altruism research in theoretical biology) remains uncommitted as to whether the underlying cause of cooperative behaviors lies in the genetic or cultural domain. Puzzle paradigm models are

often claimed to capture both genetic and cultural evolutionary processes. Here are some representative quotes:

> "Cost and benefit are measured in terms of fitness. Reproduction can be genetic or cultural." (Nowak 2006).

> "In evolutionary game theory it is not assumed that players are rational but only that successful strategies spread—by being inherited, for instance, or copied through imitation or learning." (Nowak & Sigmund 2005).

> "This score determines the player's success in the subsequent imitation/reproduction stage, either through replication and displacement or through neighbors imitating and adopting a more successful strategy. Note that the two interpretations reduce to the same dynamics." (Brandt et al. 2003).

> "It would be possible to construct an otherwise similar genetic model in which natural selection played the same role that payoff biased imitation plays in the present model" (Boyd et al. 2003).

Is it possible that the same model can capture both a genetic process and a cultural process? In theory, yes, but only under extremely strong assumptions: It requires that the cultural process and the genetic process act on the *same set of strategies* and that replication is based on the *same payoffs* in both processes. Under these conditions, certain imitative learning processes and a certain simple genetic replication process lead to identical dynamics (Schuster & Sigmund 1983; Hofbauer & Sigmund 2003). However, in general there is no reason to believe that cultural and genetic evolution act on the same strategies, nor that replication is based on the same payoffs in the two processes. Indeed, some puzzle paradigm researchers have themselves, in other work, pointed out that cultural transmission "can favor

different phenotypic variants than would be favored if the trait was genetically transmitted" (Boyd & Richerson 1985, p. 183).

We shall return to other issues with the cultural interpretation in section 4.4. Our focus here is that if it is not clear whether models describe genetic or cultural evolution it is even more difficult to say which behaviors the definition of cooperation apply to. It is the replication process, cultural or genetic, that defines what fitness is. Behaviors may have different payoff consequences depending on whether payoffs represent genetic fitness or cultural fitness. Hence, a behavior that does not count as cooperation in one process may count as cooperation in the other process.

4.1.2 The strategies cannot reasonably be study in isolation

In puzzle paradigm models the replicating units are strategies in a meta-game. A meta-game strategy includes, among other things, a rule about what strategy should be played in a particular round of a giving game. For a given puzzle paradigm meta-game we can try to infer an actual phenotype which corresponds to the strategies of that meta-game and then ask if that phenotype can reasonably be studied in isolation. In particular, behaviors in this meta-game must have a negligible correlation with behaviors in other strategic situations of fitness relevance. If this is not the case, the analysis must be extended to an even more complex meta-meta-game with an increasingly complex set of strategies.

Recall that giving games and cooperative strategies in such games are defined in terms of payoffs. This makes isolation highly unlikely. After all, the only difference between playing a giving game and other games, say a coordination game, is a small change in the payoffs of the game. Isolation would therefore require that the same genes that have broad-reaching systematic effects across a perceptually diverse set of giving games must cease to have systematic effects across a potentially perceptually similar set of coordination games. In

other words, a gene is required to impact the willingness to "cooperate" when the payoff structure is such that there is a net cost to self and a greater net benefit to others (as in a giving game), but *not* impact willingness for the same behavior in circumstances where the payoff structure is such that the relationship of these payoffs is contingent upon the behaviors of other players (as in a coordination game). This seems highly unlikely given the evidence for correlations of behavior across distinct strategic situations (e.g., Yamagishi et al. 2012).

4.2 The fitness function is not derived from interaction between the phenotype and relevant ecological factors

The next critical feature of a successful phenotypic gambit is a clear hypothesis concerning which ecological factors might explain variation in the phenotype of interest. Section 3.4.2 demonstrated how successful applications of the phenotypic gambit focus on concrete and straightforward life history factors. Puzzle paradigm papers offer up a variety of more abstract hypotheses to the effect that if the giving game is iterated within a meta-game, and the possibility of targeted punishment contingent upon behavior in the giving game is also a possibility within the meta-game, then cooperation can emerge. Importantly, punishment too is defined in terms of payoffs, as a behavior which reduces the payoff of the punished party, typically at a cost to the punisher. In the case of costly punishment, simply iterating rounds of the giving game followed by rounds of costly targeted punishment does not allow for the emergence and maintenance of cooperation. Our canonical puzzle paradigm papers go on to hypothesize that some additional factors may explain the evolution of cooperation. These factors include inter-group conflict, which creates the possibility for group selection (Boyd et al. 2003); reputation, which creates the possibility for indirect reciprocity (Nowak & Sigmund 2005); optional participation in cooperative endeavors (Hauert et al. 2007); economies of scale in either punishment or the giving game (Boyd et al.

2010); and the requisite communication and coordination of these various features. On a general level, the hypotheses that human cooperative behavior is affected by these additional factors are reasonable. However, the problem remains that in these models the core ecological factor – punishment – suffers from being abstractly defined in terms of payoffs instead of actual behavior.

To derive a fitness function, payoffs in the meta-game are simply calculated from the payoffs of the simple giving game and the payoffs of punishment. Herein lies a critical difference to the successful applications of the phenotypic gambit in section 3.4. Those applications of the phenotypic gambit used simplifying assumptions about life history to derive the fitness payoffs for concrete behaviors. No such connection is made in the puzzle paradigm, because the model behaviors are already defined in terms of payoffs. Simply put, the definition of cooperative behavior in terms of assumed payoffs creates a closed theoretical loop which prevents the connection of actual behaviors and fitness.

4.3 Puzzle paradigm models do not yield interesting predictions that can be tested in empirical work

With the fitness function in hand a researcher employing the phenotypic gambit is empowered to make some predictions concerning the variation in the phenotype of interest with respect to other potentially relevant ecological factors. For instance, recall that Hamilton's (1967) analysis yielded the prediction that female bias in the sex ratio at birth will increase with the degree to which male offspring compete and breed with their siblings. We have seen no such exciting and directly testable prediction about the world come out of the puzzle paradigm. The closest to predictions that our four sample papers get are general statements such as:

> "people should be less inclined to pay fixed than variable punishment costs" (Boyd et al. 2003)

> "if the joint enterprise is optional, cooperation backed by punishment is more likely than if the joint enterprise is obligatory." (Hauert et al. 2007)

> "the model predicts that only some individuals will engage in punishment." (Boyd et al. 2010)

The lack of exciting predictions is to be expected given the closed theoretical loop of the puzzle paradigm (section 4.2). The theory never links concrete ecological factors to concrete behaviors.

4.4 Shifting to cultural evolution does not salvage the puzzle paradigm

In section 4.1.1 we discussed that puzzle paradigm papers tend to be ambiguous as to whether the underlying replication process driving cooperative behavior is cultural or genetic. We then went on to largely ignore cultural imitation processes, critiquing the puzzle paradigm primarily under a genetic interpretation. As a final point we want to emphasize that the puzzle paradigm is not salvaged by a shift to a cultural interpretation. There are two senses in which such a simple shift is unjustified. To begin with, it would be wrong to entirely abandon the genetic level. There are many interesting questions concerning the genetic evolutionary history of human behavioral plasticity itself, as well as of the various more general cognitive capacities that enable culture and cooperation (Heyes 2012; Tomasello et al. 2005).

An even stronger reason not to shift the evolutionary puzzle of human cooperation to cultural evolution is that the shift is not valid. Evolution at the cultural level is different from evolution at the genetic level with the fundamental concepts of the latter, such as replication and fitness, failing to carry over to culture in any well-defined way (Gabora 2011; Henrich et

al. 2008). Whereas genetic fitness can be roughly measured by looking at individuals and counting their surviving descendants some time later, there is no corresponding way to measure cultural fitness by looking at an individual and counting how many others have copied her behavior. Although the term "payoff biased imitation" and the payoff based replication it implies may sound a reasonable thing, its application to puzzle paradigm models of human cooperation requires several strong assumptions to be met. One requirement is that people must be able to observe other individuals' strategies in the focal repeated game. However, it is generally not possible to infer complex strategies in repeated games from observed behavior, because the same observed behavior can be generated by many different strategies. Thus, already this fundamental requirement is unlikely to be met. The following sections discuss two additional assumptions that are necessary but unlikely to hold.

4.4.1 Imitation of someone else's strategy must be the dominant mode of behavioral change

If the strategies in puzzle paradigm models are to be interpreted as cultural replicators, they are learned and possibly relearned many times over a lifetime. But learning is not necessarily imitation. Adoption of a behavior can occur through many other pathways. Learning in any one situation may spill over to other situations; e.g., it is conceivable that someone who goes through a vaccination program might become more inclined to give blood because of the contextual similarity between the two situations. An individual may also learn general rules such as "always be considerate of others", which would clearly affect behavior in many situations. Thus, a behavioral change in one situation may be brought about by cultural learning in several strategically distinct situations, and at different levels of generality. This stands in stark contrast with the assumption that behavior in a giving game situation changes mainly due to individuals adopting a new meta-game strategy in a unique meta-game.

Or consider the impact on giving behavior of the various identities of individuals, such as their personal identity, familial identity, and community identity. Aaker and Akutsu (2009) discuss how different identities may range from specific to broad, may be activated in different contexts, may interact with emotions, may shift over the life-span, may be reinforced or change due to one's own choices to give or not, and may generally color the way in which one sees the world. See also Weber et al. (2004). It is difficult to see how the impact of context-activated identities on giving can be reconciled with the notion of imitation of meta-game strategies.

4.4.2 People must be able to observe other individuals' total payoffs in the focal repeated game and preferentially imitate high-payoff strategies

So far we have argued that it is unlikely that meta-game strategies can be observed so that they can be imitated. Moreover, we have argued that even if they could be observed, it is unlikely that behavioral change is mainly driven by imitation of such strategies. Our last point is that even if meta-game strategies are imitated, it is unlikely that people base such imitation on accurate observations of cultural fitness payoffs that individuals accrue.

The raw payoffs of behaviors come in a variety of currencies (e.g., resources, effort, pain and pleasure, and reputation). Preferences for these currencies vary. To the extent that individuals at all make decisions by adding up these incommensurable terms to a single total payoff, they will surely do so in idiosyncratic and highly subjective ways. These subjective total payoffs are not observable to others. To the extent that some component of the total payoff is observable (e.g., an individual's wealth or level of happiness), this payoff component will be subject to contributions from other parts of the individual's life as well. It seems quite impossible for others to correctly estimate what proportion of an individual's

wealth or happiness, say, should be attributed to the individual's behavior in the focal repeated game.

5. The evolutionary puzzle of human cooperation is a red herring

In section 4 we used a comparison with the phenotypic gambit in evolutionary biology to examine the puzzle paradigm for studying human cooperation. The comparison highlighted several weaknesses in puzzle paradigm research. Here we shall argue that very premise of the paradigm – that human cooperation is an evolutionary puzzle – is questionable.

To start with, it is questionable which specific puzzle the paradigm is meant to address. On the one hand, all the papers we have quoted from include various statements indicating that the primary variation of interest is inter-species variation with humans as outliers (e.g., "Unlike any other species, humans cooperate with non-kin in large groups", Boyd et al. 2003). On the other hand, other statements made in the same papers explicitly refer to within-species variation instead (see quotes in section 4.3). Thus, despite the references to human uniqueness we believe puzzle paradigm research means to address the within-species puzzle of why altruistic behavior has not been selected against. It is questionable to frame this question as an evolutionary puzzle of human cooperation to be solved by the identification of the correct meta-game. To do so requires a number of tacit assumptions:

1. The expression of altruistic behavior is the output of identifiable strategies that serve as replicators in an evolutionary process. (Otherwise it cannot be selected against.)
2. For these strategies, and the altruistic behaviors they result in, there are measurable fitness payoffs such that the altruistic behaviors have negative payoff but the strategy as a whole has not. (Otherwise there is no puzzle and no solution, respectively.)

3. The evolutionary process with respect to these strategies has reached an equilibrium. (Otherwise selection may just be ongoing.)
4. The kind of cooperation at which humans excel amounts, in fact, to altruistic behavior. (Otherwise the puzzle is not about human cooperation.)

With respect to the last point, it is certainly not clear that human cooperation is characterized by altruistic behavior. In a book devoted to human cooperation, Argyle (1991) explicitly rejects the notion of cooperation as a game theoretic strategy and instead defines it as "acting together in a coordinated way at work, leisure or in social relationships, in the pursuit of shared goals, the enjoyment of the joint activity, or simply furthering relationship". Assuming the goal is restricted to understand human altruism and not cooperation in this wider sense, we still believe it is more fruitful to question the first three assumptions than to take them for granted. Questioning these assumptions points to a number of intriguing empirical research questions. These have the potential to result in genuine knowledge and inform future attempts at theoretical modeling.

5.1 Can behavior be captured as the output of replicating strategies?

At the heart of game theoretic models lies the assumption of a certain set of strategies. As illustrated by puzzle paradigm models, strategies may be simple ("in situation A do X") or of greater complexity (e.g., "in situation A do X if C holds" or "in situation A do X and in situation B do Y"). It is a genuine challenge to assess if this concept of strategy can capture how people actually behave and, if so, what strategies are actually used. There are two main reasons why this is challenging. First, observations of behavior may be consistent with any number of different strategies (e.g., an observation of behavior X in situation A would be consistent with any of the above three examples of strategies). Second, there is no

unambiguous way to categorize situations or behaviors in the first place. In other words, given that no two situations or behaviors are ever perfectly identical, it may not be clear which should be taken as instances of A and X. Somehow these difficulties must be faced if a given game theory model is ever to convince those who do not a priori believe in the validity of its strategy set. The notion of strategies as units of replication suggests one way in which their complexity in terms of scope should be investigated, namely, analysis of how clusters of behaviors tend to vary together between individuals and change together within individuals. Such clusters are candidates for units of replication.

5.2 What are the payoffs?

What is the actual payoff, to self and others, of various strategies? This question has several layers of complexity. One problem is that it is not clear what type of payoff to measure. It is possible, though difficult, to measure both genetically relevant payoff (number of surviving children or grandchildren) and various other payoff currencies that people may care about, such as health, wealth, and life satisfaction. A more daunting problem is how to assess the link between these payoffs and the replication of specific strategies (assuming these have been identified, see section 5.1). This is comparable to the notoriously difficult problem of accurately assessing health risks and benefits of behaviors such as consumption of certain food or drink. Perhaps research methods from that field can be adapted to behaviors related to cooperation.

5.3 Are altruistic behaviors at equilibrium?

According to the most recent official Swedish statistics (https://geblod.nu/fakta/blodgivningsstatistik, accessed 4th July 2016), the number of blood donors in Sweden has declined every year between 2010 and 2015 The total decline in just five years is almost 12%, despite the population growing by almost 5% in the same period.

These data suggest that blood donations in Sweden are not at equilibrium. Before searching for equilibria in models, it seems worthwhile to study actual empirical trends in behavior and search for their causes.

5.4 Conclusion

In this essay we have criticized an approach to human cooperative behavior where evolutionary explanations are proposed and studied with mathematical models. There have been previous critiques of misapplication of evolutionary theory to human cooperation (e.g., El Mouden et al. 2012; Scott-Phillips et al. 2011; West et al. 2007; West et al. 2011). We believe that these critiques do not go far enough. Ours is more in the spirit of Arnold's (2013) recent critique of simulations of the repeated prisoner's dilemma to study the evolution of cooperation in that it questions the scientific value of an entire paradigm.

Let us emphasize that we, the authors, love both mathematics and evolutionary theory. We believe that mathematical modeling of evolutionary processes is sometimes an extremely useful research tool—but only sometimes. It is unlikely to be useful if the modeling paradigm is chosen without careful consideration of whether it fits the phenomenon one wants to study. It is fine to study models of payoff based replication of giving game strategies for their own sake. But students of human behavior should choose that particular modeling paradigm only if it is able to capture what is really going on in human behavior.

References

1. Aaker, J.L. & Akutsu, S. (2009) Why do people give? The role of identity in giving. Journal of Consumer Psychology 19:267–270.
2. Andrews, P., Gangestad, S. & Matthews, D. (2002) Adaptionism – how to carry out an adaptionist program. Behavioral and Brain Sciences 25:489–553.
3. Argyle, M. (1991) Cooperation: The Basis of Sociability. Routledge.

4. Arnold, E. (2013). Simulation models of the evolution of cooperation as proofs of logical possibilities. How useful are they?. Ethics & Politics, 2, 101-138.
5. Bowles, S. & Gintis, H. (2002) Behavioural science: homo reciprocans. Nature 415 (6868):125–127.
6. Boyd, R., Gintis, H. & Bowles, S. (2010) Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science 328 (5978):617–620.
7. Boyd, R., Gintis, H., Bowles, S. & Richerson, P. (2003) The evolution of altruistic punishment. Proceedings of the National Academy of Sciences 100 (6):3531–3535.
8. Boyd, R. & Richerson, P. (1988) The evolution of reciprocity in sizable groups. Journal of Theoretical Biology 132 (3):337–356.
9. Brandt, H., Hauert, C. & Sigmund, K. (2003) Punishment and reputation in spatial public goods games. Proceedings of the Royal Society of London. Series B: Biological Sciences 270 (1519):1099–1104.
10. Dawkins, R. (1999) The extended phenotype: The long reach of the gene. Oxford University Press, USA.
11. El Mouden, C., Burton-Chellew, M., Gardner, A. & West, S.A. (2012) What do humans maximise? In: Evolution and Rationality, ed. S. Okasha & K. Binmore, [23-49]. Cambridge University Press.
12. Fisher, R., 1930. The Theory of Natural Selection. Oxford University Press, London.
13. Fletcher, J. A., & Doebeli, M. (2006). How altruism evolves: assortment and synergy. Journal of Evolutionary Biology, 19(5): 1389-1393.
14. Gabora, L. (2011) Five clarifications about cultural evolution. Journal of Cognition and Culture 11:61– 83.
15. Gardner A. (2013) Ultimate explanations concern the adaptive rationale for organism design. Biology and Philosophy 28:787–791.

16. Gould, S. (1982) Darwinism and the expansion of evolutionary theory. Science 216 (4544):380–387.
17. Gould, S. & Lewontin, R. (1979) The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London Series B. Biological Sciences 205 (1161): 581–598.
18. Grafen, A., 1984. Natural selection, kin selection and group selection. In: Behavioural ecology: an evolutionary approach 2, eds. J. R. Krebs & N. B. Davies, [62-84]. Cambridge University Press.
19. Hamilton, W.D. (1963) The evolution of altruistic behavior. The American Naturalist 97 (896): 354–356
20. Hamilton, W. D. (1967) Extraordinary sex ratios. Science 156 (3774): 477–488.
21. Hauert, C., Traulsen, A., Brandt, H., Nowak, M. & Sigmund, K. (2007) Via freedom to coercion: the emergence of costly punishment. Science 316 (5833): 1905–1907.
22. Henrich, J., Boyd, R., & Richerson, P. J. (2008) Five misunderstandings about cultural evolution. Human Nature, 19 (2):119–137.
23. Heyes, C. (2012) New thinking: the evolution of human cognition. Philosophical Transactions of the Royal Society B: Biological Sciences 367 (1599):2091–2096.
24. Higginson, A., McNamara, J. & Houston, A. (2012) The starvation-predation trade-off predicts trends in body size, muscularity, and adiposity between and within taxa. The American Naturalist, 179 (3): 338–350.
25. Hofbauer, J. & Sigmund, K. (2003) Evolutionary game dynamics. Bulletin of the American Mathematical Society 40 (4): 479–519.
26. Kokko, H. & Jennions, M. D. (2008). Parental investment, sexual selection and sex ratios. Journal of Evolutionary Biology 21 (4): 919–948.

27. Lewontin, R. (1979) Sociobiology as an adaptationist program. Behavioral Science 24 (1): 5–14.
28. Nowak, M., 2006. Five rules for the evolution of cooperation. Science 314 (5805): 1560–1563.
29. Nowak, M. A. & Sigmund, K. (2005) Evolution of indirect reciprocity. Nature 437 (7063): 1291–1298.
30. Pen, I., & Taylor, P. D. (2005) Modelling information exchange in worker–queen conflict over sex allocation. Proceedings of the Royal Society B: Biological Sciences, 272 (1579): 2403–2408.
31. Schuster, P. & Sigmund, K. (1983) Replicator dynamics. Journal of Theoretical Biology 100 (3), 533–538.
32. Scott-Phillips, T., Dickins, T. & West, S. (2011) Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. Perspectives on Psychological Science 6 (1), 38–47.
33. Smith, J. (1978) Optimization theory in evolution. Annual Review of Ecology and Systematics 9, 31–56.
34. Smith, J. & Price, G. (1973) The logic of animal conflict. Nature 246: 15–18.
35. Taylor, P. & Jonker, L. (1978) Evolutionary stable strategies and game dynamics. Mathematical Biosciences 40 (1), 145–156.
36. Taylor, P. D. (1979) Evolutionarily stable strategies with two types of player. Journal of Applied Probability 16: 76–83.
37. Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. (2005) Understanding and sharing intentions: The origins of cultural cognition. Behavioral and Brain Sciences, 28 (5): 675–690.

38. van Veelen, M., García, J., Rand, D. G. & Nowak, M. A. (2012) Direct reciprocity in structured populations. Proceedings of the National Academy of Sciences 109 (25):9929–9934.
39. Weber, J.M., Kopelman, S. & Messick, D.M. (2004) A conceptual review of decision making in social dilemmas: Applying a logic of appropriateness. Personality and Social Psychology Review 8:281–307.
40. West, S., El Mouden, C. & Gardner, A. (2011) Sixteen common misconceptions
41. about the evolution of cooperation in humans. Evolution and Human Behavior 32 (4):231–262.
42. West, S., Griffin, A. & Gardner, A. (2007) Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. Journal of Evolutionary Biology 20 (2):415–432.
43. Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H. & Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. Proceedings of the National Academy of Sciences 109 (50):20364–20368.