Akdeniz, Aslihan; Graser, Christopher; van Veelen, Matthijs

Working Paper

Homo Moralis and regular altruists II

Tinbergen Institute Discussion Paper, No. TI 2023-025/I

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

TI 2023-025/I
Tinbergen Institute Discussion Paper

Homo Moralis and regular altruists II

Aslihan Akdeniz[1]
Christopher Graser[1,2]
Matthijs van Veelen[1]

1 University of Amsterdam and Tinbergen Institute

2 Dana-Farber Cancer Institute

Homo Moralis and regular altruists II

Aslıhan Akdeniz[1,2], Christopher Graser[1,2,3], and Matthijs van Veelen[1,2]

[1]University of Amsterdam, The Netherlands.

[2]Tinbergen Institute, The Netherlands.

[3]Dana-Farber Cancer Institute, Boston, USA.

May 2, 2023

Abstract

Alger and Weibull (2013) ask the question whether a combination of assortative matching and incomplete information leads to the evolution of moral or altruistic preferences. Their central result states that Homo Hamiltonenis – a type that has moral preferences with a morality parameter equal to the level of assortment – is evolutionarily stable, while preferences that lead to different behaviour are unstable. Together with their claim that altruistic and moral preferences differ sharply, this suggests that moral preferences tend to beat altruistic ones in evolutionary competition. We show that this is not true. First of all, we show that there is a loophole in the definition of evolutionary stability, allowing for Homo Hamiltonensis to satisfy the definition when the set of equilibria is empty, and their equilibrium behaviour is not determined. If we try to close this loophole, by allowing for mixing, or by allowing for asymmetric equilibria, we find that there are two options. With the first approach, the differences in behaviour between Homo Hamiltonensis and regular altruists can be substantial, but as soon as the difference appears, Homo Hamiltonensis can be invaded, and regular altruists win in direct competition. With the second way of allowing for mixing, or coordination on asymmetric equilibria, Homo Hamiltonensis cannot be invaded, but then the difference in behaviour all but disappears, as all equilibria between Homo Hamiltonensis are also equilibria between regular altruists.

Altruism, morality, evolution, assortment, incomplete information

1 Introduction

In 1964, Bill Hamilton published a paper in which he presented what is now known as *Hamilton's rule*. This rule states that altruism will evolve if $rb > c$, where b is the fitness benefit to the recipient, c is the fitness cost to the donor, and r is the genetic relatedness between them. The idea that relatedness can breed altruism was a breakthrough, that is reflected in an enormous amount of research, as well as many popular science books, including the most popular of popular science books, *The Selfish Gene* by Richard Dawkins (1976).

In a more recent paper, Alger and Weibull (2013) suggest that relatedness – or assortment[1] – can also lead to something other than altruism. Altruism can be described by a utility function that puts a positive weight on the fitness, or material payoff, of the other – besides a positive weight on the fitness, or material payoff, of oneself. The preference that Alger and Weibull (2013) focus on instead maximizes a convex combination of one's own material interests, and a hypothetical payoff. This hypothetical payoff is the material payoff that one would get if both were to play the strategy that one plays oneself. A type that maximizes such a convex combination is referred to by the authors as *Homo Moralis*. A Homo Moralis that puts a weight on this hypothetical payoff that is equal to the assortment parameter of the model is referred to by the authors as *Homo Hamiltonensis*.

The aim of their paper is to determine which one of those preferences evolution would select; altruists or Homo Moralis. In the words of the authors (page 2270):

However, this literature is silent as to whether either altruistic or moral preferences would, in fact, arise if evolution were to operate on preferences — as a way for nature to delegate the choices of concrete actions to the individual in any given situation. It is our goal to fill this gap.

Their central result points to Homo Hamiltonensis as the winner of this evolutionary contest. In their model of preference evolution (that combines assortative matching with imperfect information about each other's preferences), and under conditions specified in the paper, *Homo Hamiltonensis* satisfies their definition of evolutionary stability. Their central result also states that preferences that lead to different behaviour are evolutionarily unstable.

In order to be able to reflect on their argument, it is good to spend some attention on two more ingredients. The first is that the method that Alger and Weibull (2013) aim to use in

[1] Relatedness can be described as assortment of genotypes due to population structure, where behaviour is transmitted genetically. Alger and Weibull (2013) assume assortment without making assumptions about the mode of transmission (which can be genetic or cultural) or what it is that generates the assortment (which could for instance be kin recognition, or a combination of local interaction and local dispersal). Kay, Keller, and Lehmann (2020) suggest that one could use the word relatedness so as to also include cultural transmission as a cause of identity by descent, while others restrict the term relatedness to genetic transmission.

their paper is the "indirect evolutionary approach". The idea of this approach is that the utility functions that players have (which types they are) determine their behavior, their behavior determines their material payoffs, and the material payoffs determine which utility functions (types) get selected.

The second ingredient is that in their paper, Alger and Weibull (2013) stress that altruism and morality, in their definition, are not the same. On page 2270, they write:

Clearly, these two motivations may give rise to different behaviors.

When they discuss the differences in more detail, on page 2293, they write:

As noted above, the preferences of homo moralis *differ sharply from altruism. We first show that, while in some situations morality and altruism lead to the same behaviour, in other situations, the contrast is stark.*

The argument in the paper therefore has three main ingredients; 1) their central result that states that Homo Moralis satisfies their definition of evolutionary stability, and types that play differently do not; 2) their claim that they use the "indirect evolutionary method"; and 3) their observation that there can be a stark contrast between the behaviour of the Homo Moralis and regular altruists.

The message of these three ingredients together is crystal clear: when assortment is combined with imperfect information, and Homo Moralis and regular altruists differ in their behaviour, we should expect morality to evolve, and not altruism. Altruism and morality sometimes lead to the same behaviour, but they also regularly do not, and if they do not, then the evolutionary competition between them is won by morality, and not by altruism, because Homo Moralis is evolutionarily stable, and preferences that make one play differently are not.

In this paper, we will show that this is not correct. That seems surprising, because this is not due to their central result not holding. It will therefore be important to determine precisely how it is possible that no examples exist of cases in which there is a "stark contrast" between regular altruists and Homo Hamiltonensis, while the latter beats the former – even though that is what the paper suggests. We will do this with an instructive example that shows that, if anything, the *opposite* conclusion (that regular altruists win the competition, if their behaviours differ) is true, and with two general results.

In search of a game in which regular altruists and Homo Hamiltonensis display different behaviour, we first encounter a situation in which Homo Hamiltonensis is determined to be

evolutionarily stable through a loophole in the definition. If the set of Bayesian Nash Equilibria, as defined by Alger and Weibull (2013), is empty, and neither the material payoffs of Homo Hamiltonensis as a resident, nor the material payoff of any possible mutant is defined, then Homo Hamiltonensis nonetheless satisfies the definition of evolutionary stability automatically. This is obviously not based on a comparison of material payoffs – which are not defined – and therefore the definition of evolutionary stability in this case does not reflect the indirect evolutionary approach.

In this example – which squarely falls in the domain that their central result applies to – only pure strategies can be used. If we try to remedy the absence of BNE by allowing for mixing, then two possibilities arise. With the first way of allowing for mixing, there is a difference in behaviour between regular altruists on the one hand, and the Homo Hamiltonensis that we get if we extend it to lotteries this way on the other (where we will call this extended version Homo Hamiltonensis 1.0). As soon as they start behaving differently, however, regular altruists start getting *higher* material payoffs than Homo Hamiltonensis 1.0, when Homo Hamiltonensis 1.0 is the resident, and the regular altruist is the mutant. Regular altruists therefore can invade, while they themselves cannot be invaded.

With the second way of allowing for mixing, leading to what we call Homo Hamiltonensis 2.0, Homo Hamiltonensis stops getting lower payoffs than regular altruists, but also the difference in behaviour disappears. Our Proposition 1 shows that this reflects something that is true generally; any equilibrium for a Homo Hamiltonensis 2.0 is also an equilibrium between regular altruists. This result also implies that as soon as there is a difference between Homo Hamiltonensis 1.0 and regular altruists, Homo Hamiltonensis 1.0 can be invaded, and typically they can be invaded by regular altruists.

We then repeat this, but instead of allowing for mixing, we allow for asymmetric equilibria (while the initial situation only allows for pure, symmetric equilibria). The results are the same as with mixing. With symmetric equilibria only, Homo Hamiltonensis and regular altruists either behave the same, or Homo Hamiltonensis is declared evolutionarily stable through a loophole in the definition. With one way of allowing for asymmetric equilibria, Homo Hamiltonensis and regular altruists display different behaviours, but Homo Hamiltonensis can be invaded, and regular altruists cannot. With the other way of allowing for asymmetric equilibria, Homo Hamiltonensis can no longer be invaded, but now the "stark contrast" between Homo Hamiltonensis and regular altruists dissipates. Proposition 2 is the counterpart of Proposition 1, with asymmetric equilibria instead of mixing, and this proposition also indicates that our example captures a general observation; it rules out the possibility that Homo Hamiltonensis both "differs sharply" from regular altruists, and beats them in evolutionary competition. Since the details of the

definitions turn out to be crucial, we repeat the model before we get to our example, and our two general results.

2 The Model

Alger and Weibull (2013) consider a population in which individuals are matched in pairs to engage in a symmetric interaction with a common strategy set X. These individuals have preferences, and the idea is that these preferences determine what the individuals that have them do. We will see that one of the problems is caused by the fact that Homo Hamiltonensis can satisfy the definition of evolutionary stability also if its equilibrium behaviour is not defined, but for now we can imagine a situation in which the preferences that a type holds do imply well-defined equilibrium behaviour.

What the individuals do in the interaction then determines their evolutionary success, according to a material payoff function $\pi(x, y)$, where $\pi : X^2 \to \mathbb{R}$. To study the evolution of preferences, Alger and Weibull (2013) consider a situation with a resident type θ and a mutant type τ, where θ and τ are preferences that individuals can have over strategy profiles; $u_\theta : X^2 \to \mathbb{R}$ and $u_\tau : X^2 \to \mathbb{R}$.

The players in the population are not matched uniformly randomly. Instead, an assortment parameter σ is introduced, and this parameter defines the probabilities with which these two types interact in the limit of vanishing mutant shares ϵ. In this limit, the resident is always matched with another resident; $\lim_{\epsilon \downarrow 0} Pr[\theta|\theta, \epsilon] = 1$. In the same limit, the mutant is matched with another mutant with probability $\lim_{\epsilon \downarrow 0} Pr[\tau|\tau, \epsilon] = \sigma$, and with a resident with probability $\lim_{\epsilon \downarrow 0} Pr[\theta|\tau, \epsilon] = 1 - \sigma$. One way to interpret this would be that, in this limit, every individual is matched with a random draw from the population with probability $1 - \sigma$, and with a copy of itself with probability σ.[2] Since a random draw at mutant frequency 0 means being matched to a resident for sure, this gives the limiting probabilities as described.

It is assumed that these individuals do not know the preferences of the individual they are matched with, but they do know what their own preferences are, and what that implies for their probabilities of being matched with either type. Also mutants know what the preferences of the resident are, residents know what the preferences of the mutant are, and both know in which shares they occur, or, in other words, both are aware of the ϵ. The choices that residents θ and mutants τ make are assumed to constitute a symmetric pure (Bayesian) Nash Equilibrium (BNE), given a population state $s = (\theta, \tau, \epsilon) \in S$, where $\theta, \tau \in \Theta$ are the resident and the mutant

[2] This interpretation does not have to be restricted to the limit; see van Veelen (2009, 2011); van Veelen, Allen, Hoffman, Simon, and Veller (2017); van Veelen (2018). Also in the examples below we will assume population structures where this interpretation extends to $\epsilon > 0$.

type, and where Θ is the set of types we are considering. This makes the set of population states $S = \Theta^2 \times (0,1)$.

Definition 1. *In any state $s = (\theta, \tau, \epsilon) \in S$, a strategy pair $(x^*, y^*) \in X^2$ is a (Bayesian) Nash Equilibrium (BNE) if*

$$x^* \in \arg\max_{x \in X} Pr[\theta|\theta,\epsilon] \cdot u_\theta(x, x^*) + Pr[\tau|\theta,\epsilon] \cdot u_\theta(x, y^*),$$
$$y^* \in \arg\max_{y \in X} Pr[\theta|\tau,\epsilon] \cdot u_\tau(y, x^*) + Pr[\tau|\tau,\epsilon] \cdot u_\tau(y, y^*).$$

The set of Bayesian Nash equilibria for state (θ, τ, ϵ) is denoted by $B^{NE}(\theta, \tau, \epsilon)$. It is important to remember that the setup of the model explicitly allows for X to be a set of pure strategies. The definition of a BNE does not allow players to mix over different elements of X, and therefore, if we choose X to be a set of pure strategies, mixed equilibria are not allowed. The set of BNE therefore can be empty, and it is also possible that the set of BNE is empty without allowing for mixing, while it would not be empty if we were to allow for mixing. The definition moreover does not allow for asymmetric equilibria - in which players would be randomly assigned to role 1 or 2, and in which either the resident, or the mutant, or both would condition their behaviour on whether they are player 1 or 2. It is also possible that the set of BNE is empty if asymmetric equilibria are not allowed for, and not empty if we do allow for asymmetric equilibria.

What is and what is not a BNE will typically depend on ϵ. Since we are interested in evolutionary stability, we want to look at what happens for small ϵ. If for small enough ϵ all BNE, as well as all conditional probabilities, change continuously as a function of ϵ, and if the limiting equilibria for $\epsilon \downarrow 0$ equal the equilibria at $\epsilon = 0$, then the equilibria in this limit will be relevant for the stability of type θ against τ. At $\epsilon = 0$ the resident only meets copies of itself, and therefore the following, simpler equations make (x^*, y^*) a symmetric pure BNE at $\epsilon = 0$.

$$x^* \in \arg\max_{x \in X} u_\theta(x, x^*) \tag{1}$$
$$y^* \in \arg\max_{y \in X} (1-\sigma) \cdot u_\tau(y, x^*) + \sigma \cdot u_\tau(y, y^*) \tag{2}$$

The average material payoffs, or fitnesses, of the different types depend on what they do, and who they are matched with. If θ-types play x and τ-types play y, then the resulting material

payoffs, or fitnesses, are

$$\Pi_\theta(x,y,\epsilon) = Pr[\theta|\theta,\epsilon]\cdot\pi(x,x) + Pr[\tau|\theta,\epsilon]\cdot\pi(x,y)$$
$$\Pi_\tau(x,y,\epsilon) = Pr[\theta|\tau,\epsilon]\cdot\pi(y,x) + Pr[\tau|\tau,\epsilon]\cdot\pi(y,y)$$

If we assume that a BNE is played, then $\Pi_\theta(x,y,\epsilon)$ will, obviously, depend on θ, but also on what τ is, and $\Pi_\tau(x,y,\epsilon)$ will, besides on τ, also depend on what θ is. This is suppressed in the notation. In case of multiple equilbria, the payoffs will also depend on which equilibrium is played.

Which strategies the resident can play in a BNE will become independent of which mutant τ we are considering at $\epsilon = 0$. If we consider a strategy profile (x^*, y^*) that is a BNE at $\epsilon = 0$ for a resident θ and a mutant τ, then we can denote their payoffs as follows[3]:

$$\Pi_\theta(x^*) = \pi(x^*,x^*)$$
$$\Pi_{\tau,\theta}(x^*,y^*) = (1-\sigma)\cdot\pi(y^*,x^*) + \sigma\cdot\pi(y^*,y^*)$$

One could call $\Pi_\theta(x^*)$ the fitness of the resident θ for x^*, and one could call $\Pi_{\tau,\theta}(x^*,y^*)$ the invasion fitness of mutant τ for (x^*,y^*) – which is assumed to be a BNE. If there is a unique x^* such that $x^* \in \arg\max_{x\in X} u_\theta(x,x^*)$, then one could call $\Pi_\theta = \Pi_\theta(x^*)$ the fitness of the resident. This fitness is then naturally independent of the mutant type. If there is moreover also a unique y^* such that $y^* \in \arg\max_{y\in X}(1-\sigma)\cdot u_\tau(y,x^*) + \sigma\cdot u_\tau(y,y^*)$, then one could call $\Pi_{\tau,\theta} = \Pi_{\tau,\theta}(x^*,y^*)$ the invasion fitness of mutant τ, and this will typically depend on the resident type θ. In the examples below, we will, among other things, calculate these for resident Homo Hamiltonensis and mutant regular altruists.

Definition 2. *A type $\theta \in \Theta$ is evolutionarily stable against a type $\tau \in \Theta$ if there exists an $\bar{\epsilon} > 0$ such that $\Pi_\theta(x^*,y^*,\epsilon) > \Pi_\tau(x^*,y^*,\epsilon)$ in all Nash equilibria (x^*,y^*) in all states $s = (\theta,\tau,\epsilon)$ with $\epsilon \in (0,\bar{\epsilon})$. A type θ is evolutionarily stable if it is evolutionarily stable against all types $\tau \neq \theta$ in Θ*

The main result in Alger and Weibull (2013) is that, under some restriction on the payoff function π, a type that they call Homo Hamiltonensis is evolutionarily stable against all types that are not behavioural alikes. Moreover, they show that all types that are not behaviourally equivalent to Homo Hamiltonensis are evolutionarily unstable, and can be invaded, provided that the set of mutants is sufficiently large, so that it includes a mutant that would choose the strategy that one would have to play to achieve this higher material payoff. Homo Hamiltonensis

[3] This notation is not in the original paper.

is a special case of what they label a Homo Moralis, and Homo Moralis has a utility function that puts positive weights on her own material payoff, and on the (hypothetical) payoff that she and the individual she is matched with would get, if both were to play the strategy that she plays herself.

Definition 3. *A Homo Moralis with morality parameter κ maximizes the following utility function:*

$$u_\kappa = (1-\kappa)\cdot\pi(x,y) + \kappa\cdot\pi(x,x)$$

A Homo Hamiltonensis is a Homo Moralis with $\kappa = \sigma$.

The restriction that is imposed on the payoff function π in the central result is that if we consider a situation where a Homo Hameltonensis plays against a copy of itself, the best response in all Nash equilibria would have to be unique. Also, π, as well as all utility functions that define the types, are assumed to be continuous.

Before we can state their central result, we need a bit more notation. For each type $\theta \in \Theta$, $\beta_\theta : X \Rightarrow X$ denotes the best-reply correspondence:

$$\beta_\theta(y) = \arg\max_{x\in X} u_\theta(x,y) \forall y \in X$$

,

Moreover, $X_\theta \subseteq X$ is the set of fixed points under β_θ,

$$X_\theta = \{x \in X : x \in \beta_\theta(x)\}$$

In particular, X_σ is the fixed-point set for Homo Hamiltonensis.

For any type $\theta \in \Theta$, let Θ_θ be the set of types τ that, as vanishingly rare mutants among residents of type θ, are behaviorally indistinguishable from the residents[4]

$$\Theta_\theta = \left\{\tau \in \Theta : \exists x \in X_\theta \text{ such that } (x,x) \in B^{NE}(\theta,\tau,0)\right\}$$

Finally, the type set Θ will be said to be rich if, for each strategy $x \in X$, there exists some type $\theta \in \Theta$ for which this strategy is strictly dominant: $u_\theta(x,y) > u_\theta(x',y) \forall x' \neq x, \forall y \in X$. Such a type θ will be said to be committed to its strategy x.

[4]There are some problems with that definition. Behavioral alikes are only required to behave alike at $\epsilon = 0$, while that does not have to imply that they also behave alike for $\epsilon \in (0,\bar{\epsilon})$ for some $\bar{\epsilon} > 0$. In Akdeniz, Graser, and van Veelen (2020) we give an example where behavioral alikes do strictly better at $\epsilon > 0$, and therefore can invade. Also it is problematic that for being a behavioral alike, it is enough to behave alike at one BNE, while there may be other BNE at which they behave differently. Being a behavioral alike excludes mutants from payoff scrutiny, while these mutants may actually do better at all other equilibria, other than the one at which it behaves the same as the resident. In this paper, however, we focus on a different problem.

If $\beta_\sigma(x)$ is a singleton for all $x \in X_\sigma$, then homo hamiltonensis is evolutionarily stable against all types $\tau \notin \Theta_\sigma$. If Θ is rich, $X_\theta \cap X_\sigma = \varnothing$, and X_θ is a singleton, then θ is evolutionarily unstable.

The central question of Alger and Weibull (2013) is whether, if evolution operates on preferences, we should expect altruistic preferences to evolve, or moral ones. The theorem states that, under some condition, Homo Hamiltonensis (which has moral preferences with a morality parameter equal to the assortment parameter) is evolutionarily stable against all types that are not behaviourally equivalent, and, also under conditions, that proper different preferences are evolutionarily unstable. This suggests that the question of the paper is settled in favour of moral preferences, as soon as moral preferences lead to behaviour that is different from the behaviour that altruistic preferences lead to. The example below, along with Propositions 1 and 2, challenges that.

3 An example

Consider the strategy set $X = [1, \infty)$ and the payoff function[5]

$$\pi(x, y) = a(x^\beta + y^\beta)^{\frac{1}{\beta}} - x^2$$

This can be seen as a continuous public goods game, where the production function of the public good is $a(x^\beta + y^\beta)^{\frac{1}{\beta}}$, and the cost of providing the input is x^2. We assume that $\beta > 0$. For $0 < \beta < 1$, this game has strategic complements. For $\beta > 1$, it has strategic substitutes. For $\beta \to \infty$, this game converges to what one could call the maximum effort game with quadratic costs; $\pi(x, y) = a \max\{x, y\} - x^2$. Please note that this choice of a strategy set and a material payoff function satisfies all the requirements needed for the central result to apply.

To make the solutions easy to read, we will make a depend on β, and choose $a = 2^{2-\frac{1}{\beta}}$. For this choice of a, if there is a pure equilibrium at $\epsilon = 0$ between a resident Homo Hamiltonensis and any mutant, or between a resident regular altruists and any mutant, it will require the Homo Hamiltonensis, or the regular altruist, to play $1 + \sigma$, regardless of the β (see Appendix 7.1; this follows directly from the first order condition, which is the same for both types). Also between a resident Homo Hamiltonensis and a mutant regular altruist, or vice versa, the only candidate for a pure BNE is for both to play $1 + \sigma$, in this case for every $\epsilon \in [0, 1]$. Whether such BNE exist, depends on β. We can recognize three cases, separated by two thresholds, $\beta_1(\sigma)$ and $\beta_2(\sigma)$. How these thresholds depend on σ is described in Appendix 7.1. The first is always the lower one.

[5] The relevant properties of the example remain the same if we choose $X = [0, \infty)$ with the same payoff function, but for calculating the mixed equilibria, it helps avoiding needless complications if we chose $X = [1, \infty)$.

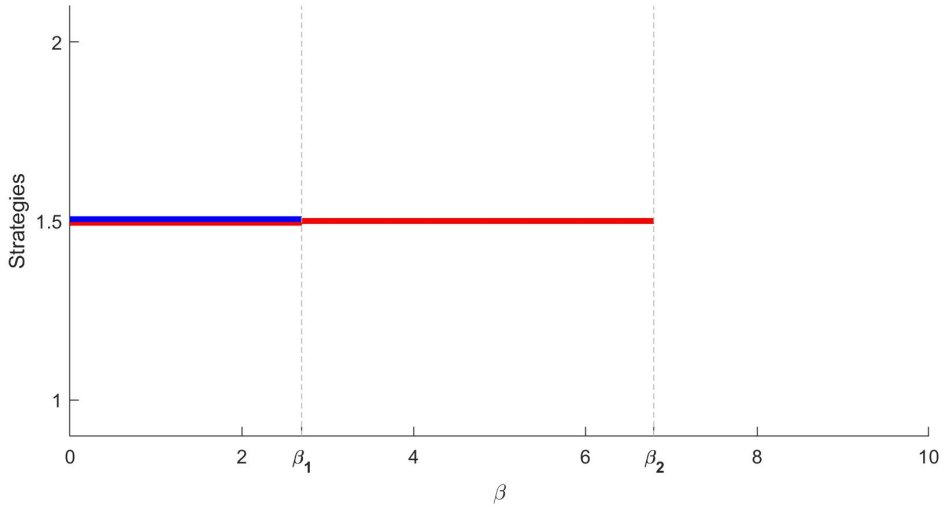

Figure 1: Behavior of a resident Homo Hamiltonensis (red) or a resident regular altruist (blue) in a Bayesian Nash equilibrium according to the definition in Alger and Weibull (2013) applied to our example, with $X = [1, \infty)$ and $\sigma = \frac{1}{2}$. At $\epsilon = 0$, the equilibrium behaviour does not depend on what type the mutant is.

Resident Homo Hamiltonensis

If we take type θ to be a Homo Hamiltonensis, then for $\beta < \beta_2(\sigma)$, the x^* that satisfies $x^* \in \arg\max_{x \in X} u_\theta(x, x^*)$ is $x^* = 1 + \sigma$, and it is the only x^* that does. This is illustrated for $\beta = 2$ and $\beta = 4$ and $\sigma = \frac{1}{2}$ in panels D and E in Fig. 2. These fixed points define the equilibrium behaviour of a resident Homo Hamiltonensis in a BNE at $\epsilon = 0$ (see equation 1). For every mutant that at $\epsilon = 0$ would play a pure strategy y^* against a resident that plays $1 + \sigma$, these strategies combine to a strategy profile (x^*, y^*) that is a BNE at $\epsilon = 0$ (see equation 2). With continuity, this extends to some interval $(0, \bar{\epsilon})$, in which a continuously changing pure strategy profile $(x^*_\epsilon, y^*_\epsilon)$, with $(x^*_0, y^*_0) = (x^*, y^*)$, would satisfy the definition of a BNE.

For $\beta > \beta_2(\sigma)$, on the other hand, there is no fixed point of equation 1 in pure strategies (see panel F in Fig. 2). That means that at $\epsilon = 0$, the set of BNE is empty for a resident Homo Hamiltonensis in combination with every possible mutant. This also extends at least to some small enough interval of positive ϵ's.

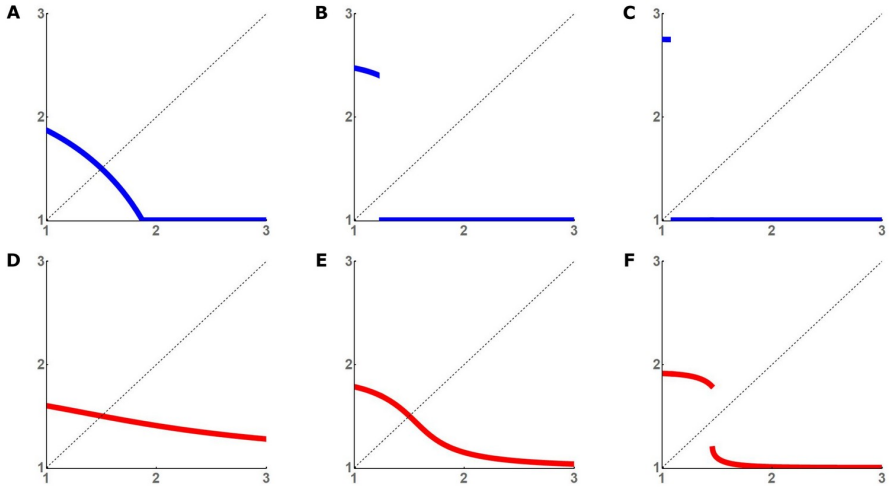

Figure 2: All panels plot $\arg\max_{x \epsilon X} u_\theta(x, x^*)$ as a function of x^*. The definition of a BNE at $\epsilon = 0$ requires that we choose x^* to be a fixed point. For the top three, type θ is a regular altruist, for the bottom three, type θ is a Homo Hamiltonensis. The β's are 2 (left), 4 (middle), and 8 (right).

A first important consequence is that for $\beta > \beta_2(\sigma)$, Homo Hamiltonensis therefore automatically satisfies the definition of evolutionary stability. The reason for this is that anything is true for all elements of the empty set (see Definition 2). In this case, it would be hard to defend that this definition accurately reflects the indirect evolutionary approach. By lack of a BNE, neither the material payoffs earned by the resident, nor the material payoffs earned by any of the mutants are defined. Satisfying the definition of evolutionary stability therefore cannot be based on comparisons of those.

Resident regular altruist

If we now take type θ to be a regular altruist, then all of these observations apply again, but now the relevant threshold is $\beta_1(\sigma)$ instead of $\beta_2(\sigma)$. This is also illustrated in Fig. 2, panel A for a β below $\beta_1(\sigma)$, and panel B and C for a β above $\beta_1(\sigma)$. Given that $\beta_1(\sigma) < \beta_2(\sigma)$, that means that there are now three regimes, visualized in Fig. 1.

Pure equilibria for both types of residents: $\beta < \beta_1(\sigma)$

For β below the lowest of the two thresholds, both a resident Homo Hamiltonensis and a resident regular altruist would play $x^* = 1 + \sigma$ at $\epsilon = 0$ against any mutant. For a mutant that plays a pure strategy at $\epsilon = 0$ against a resident playing $x^* = 1 + \sigma$, a BNE exists, both with a resident Homo Hamiltonensis, and with a resident regular altruist. No such mutant can invade either of

the two, as they would get lower material payoffs against both of them (unless the mutant also plays $y^* = 1 + \sigma$, in which case they get the same payoffs).

Pure equilibria for neither resident: $\beta > \beta_2(\sigma)$

For β above the highest of the two thresholds, both for a resident Homo Hamiltonensis, and for a resident regular altruist, the set of BNE is empty for every mutant. Both of them therefore automatically get labeled evolutionarily stable, and for both of them, this does not reflect the indirect evolutionary approach.

Pure equilibria for resident Homo Hamiltonensis, but not for resident regular altruists: $\beta_1(\sigma) < \beta < \beta_2(\sigma)$

In the middle interval, Homo Hamiltonesis as a resident plays pure strategy $x^* = 1 + \sigma$, while the equilibrium behaviour of resident regular altruists is undefined, as they have empty sets of BNE against all mutants. Moreover, for a resident Homo Hamiltonensis, mutants that would mix against them are not considered, when determining whether or not it satisfies the definition of evolutionary stability. That includes regular altruists, that, as mutants, would also mix against a resident Homo Hamiltonensis, if they could. This points to a second way in which the definition does not reflect the indirect evolutionary approach for this example. The material payoffs between a resident Homo Hamiltonensis and a mutant regular altruist are not defined, and yet, according to the definition, a resident Homo Hamiltonensis is evolutionary stable against a mutant regular altruist – even if, as we will see below, these would get higher material payoffs as mutants if they were allowed to mix.

Homo Hamiltonensis never outperforms regular altruists

It is important to note that in this example, there is no β for which a mutant regular altruist ever gets material payoffs that are lower than the material payoffs of a resident Homo Hamiltonensis. In this example, they are either playing the same strategy, and get the same material payoffs (for low β's), or the behaviour of a regular altruists as a mutant is not defined (for intermediate β's), or the behaviour of neither of them is defined (for high β's). It is also important to note that without mixing, this is not just a property of a far-fetched example. Because the first order condition for a fixed point is the same for a Homo Hamiltonensis and a regular altruists, the difference between them can only come from the second order condition, or from local maxima not being global maxima (see Appendix 7.1). That means that the typical way for interior fixed points of the two types not to coincide, is for one of them not to have a fixed point. Finally it is important to keep in mind that this example satisfies the requirements of the central result

in Alger and Weibull (2013), while it is clear that Homo Hamiltonensis satisfies the definition of evolutionary stability through a loophole for $\beta > \beta_1(\sigma)$.

4 Allowing for mixing

4.1 Method 1

The failure of the definition of evolutionary stability to reflect the indirect evolutionary approach for this example is caused by the fact that the material payoffs needed for comparisons are not defined, when the set of BNE's is empty. A natural thing to consider, in order to avoid this, is to allow for mixed equilibria. There are two ways in which one can do this. In this subsection, we will discuss the first way of including mixing. This produces non-empty sets of BNE, and equilibrium behaviour in which Homo Hamiltonensis and regular altruists (both extended to include lotteries in their domain) choose different strategies. Between the two, whenever their equilibrium behaviours differ, it is the regular altruists that get the higher material payoffs.

After this subsection, we will discuss a second way of including mixed equilibria, but we will do this more generally before applying it to this example. With this second approach, there are also non-empty sets of BNE, but now the difference in equilibrium behaviour between Homo Hamiltonensis and regular altruists disappears again.

A definition of a BNE that allows for mixed equilibria

With the first approach to allowing for mixed equilibria, we simply change the definition of a BNE so that it also allows for mixed equilibria. We do this by replacing pure strategies x^* and y^* by probability measures μ^* and ν^*

Definition 4. *In any state $s = (\theta, \tau, \epsilon) \in S$, a strategy pair (μ^*, ν^*) is a (Bayesian) Nash Equilibrium (BNE) if the following holds for all x^* in the support of μ^* and all y^* in the support of ν^**

$$x^* \in \arg\max_{x \in X} Pr[\theta|\theta,\epsilon] \cdot \int u_\theta(x,z)\, \mathrm{d}\mu^*(z) + Pr[\tau|\theta,\epsilon] \cdot \int u_\theta(x,z)\, \mathrm{d}\nu^*(z)$$
$$y^* \in \arg\max_{y \in X} Pr[\theta|\tau,\epsilon] \cdot \int u_\tau(y,z)\, \mathrm{d}\mu^*(z) + Pr[\tau|\tau,\epsilon] \cdot \int u_\tau(y,z)\, \mathrm{d}\nu^*(z)$$

Changing from Definition 1 to 4 implies that we have extended the domain of the utility functions of Homo Hamiltonensis and regular altruists, so that they also apply to lotteries, and that we have done so by assuming expected utility. For a regular altruist, that is more or less

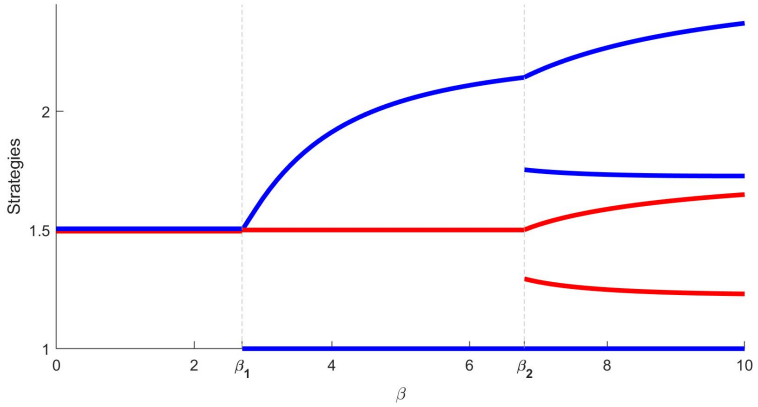

Figure 3: Equilibrium behavior of a resident Homo Hamiltonensis 1.0 (red) or a resident regular altruist (blue) in Bayesian Nash equilibria according to the definition in Alger and Weibull (2013), but extended to allow for mixed equilibria, again for $\sigma = \frac{1}{2}$. At $\epsilon = 0$, the equilibrium behaviour does not depend on what type the mutant is. In the limit of $\beta \downarrow \beta_1$, the probability on the lower value for the regular altruist is 0, and so is the probability on the middle value for regular altruists, and on the lower value for Homo Hamiltonensis 1.0 in the limit of $\beta \downarrow \beta_2$

the only reasonable option, but for Homo Hamiltonensis, as we will see below, there are also other ways to extend the domain of the utility function. Using expected utility however is not unnatural as a first step in exploring how we can make non-empty sets of equilibria with diverging behaviours. In order to have a relatively short way of referring to the version of Homo Hamiltonensis that we get by taking the Homo Hamiltonensis from a setting with pure strategies only, and extending it to cover lotteries by assuming expected utility, we will call this extended version Homo Hamiltonensis 1.0. This also helps distinguishing it from the alternative way of extending, that we will refer to as Homo Hamiltonensis 2.0. For regular altruists, the only relevant extension is to combine it with expected utility.

Nothing changes for $\beta < \beta_1(\sigma)$

If we do indeed allow for mixing in this way, then for $0 < \beta < \beta_1(\sigma)$, nothing changes; a BNE for a resident Homo Hamiltonensis 1.0 would require it to play $x^* = 1 + \sigma$, and the same is true for a resident regular altruist. Also between a resident Homo Hamiltonensis and a mutant regular altruist, the unique BNE is still $(x^*, y^*) = (1 + \sigma, 1 + \sigma)$ for every $\epsilon \in [0, 1]$.

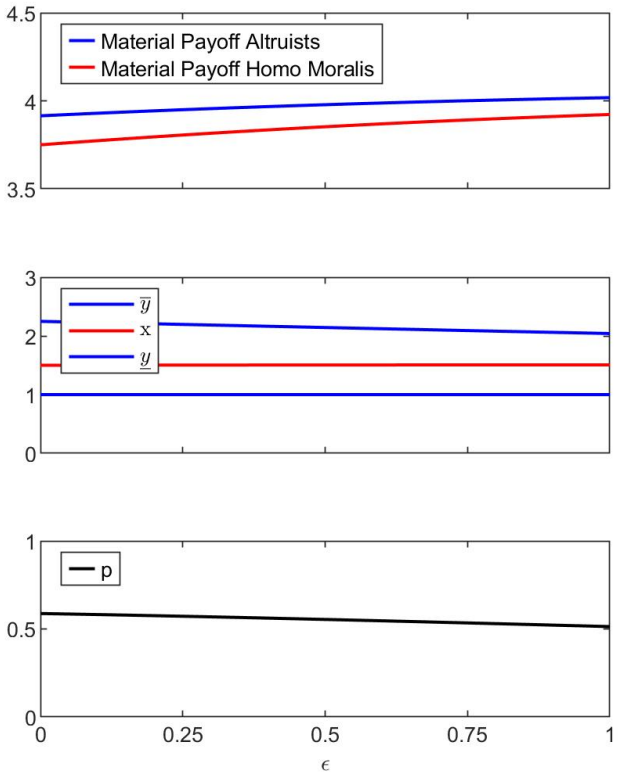

Figure 4: For $\sigma = \frac{1}{2}$ and $\beta = 5$, this represents the properties of the BNE between a resident Homo Hamiltonensis 1.0 and a mutant regular altruist for all $\epsilon \in [0, 1]$. The middle panel shows the equilibrium strategy of the resident Homo Hamiltonensis 1.0, which is $1 + \sigma = \frac{3}{2}$ for all ϵ, and the pure strategies that a regular altruists is mixing over, as functions of ϵ. The bottom panel shows the probability with which the regular altruist plays the lower of the two pure strategies. The top panel shows the material payoffs of the two types in such populations.

Altruists start mixing, and earn higher material payoffs for $\beta_1(\sigma) < \beta < \beta_2(\sigma)$

For $\beta_1(\sigma) < \beta < \beta_2(\sigma)$, a resident Homo Hamiltonensis 1.0 would still play a pure strategy $1 + \sigma$ at $\epsilon = 0$ for every mutant, but a resident regular altruist would play a mix. Moreover, between a resident Homo Hamiltonensis and a mutant regular altruist, the mutant altruist would mix. An important observation here is that a resident Homo Hamiltonensis 1.0 can be invaded (by a mutant regular altruist, among others), while a resident regular altruist cannot be invaded by anything. In direct competition, regular altruist outperform Homo Hamiltonensis 1.0 at all shares ϵ and for all β's in this interval. As an example, Fig. 4 plots all relevant properties of the BNE, allowing for mixing, for $\sigma = \frac{1}{2}$ and $\beta = 5$, as functions of ϵ; it plots the pure strategy

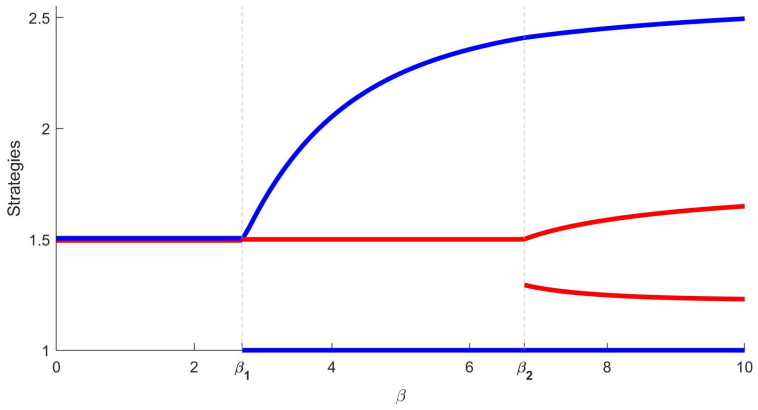

Figure 5: Equilibrium behaviour between a resident Homo Hamiltonensis 1.0 (red) and a mutant regular altruist (blue) at $\epsilon = 0$. The mutant regular altruist always outperforms the resident Homo Hamiltonensis 1.0 at $\epsilon = 0$ (see Fig. 6).

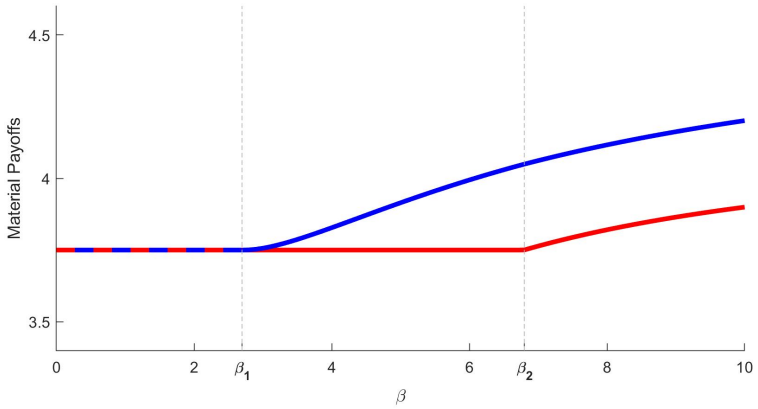

Figure 6: Material payoffs for the BNE, depicted in Fig. 5 between a resident Homo Hamiltonensis 1.0 (red) and a mutant regular altruist (blue) at $\epsilon = 0$.

chosen by the Homo Hamiltonensis 1.0; the two pure strategies that regular altruists are mixing over; the equilibrium probability on the first; and both types' material payoffs, all for $\sigma = \frac{1}{2}$

and $\beta = 5$. There, we can see that not only can regular altruists invade (the material payoffs of regular altruists are higher than the material payoffs of Homo Hamiltonensis 1.0 at $\epsilon = 0$), and not only can Homo Hamiltonensis 1.0 *not* invade regular altruists (the material payoffs of regular altruists is also higher than the material payoffs of Homo Hamiltonensis 1.0 at $\epsilon = 1$), but regular altruists actually have higher payoffs for every mutant share $\epsilon \in [0, 1]$. This is representative for all σ and all β between $\beta_1(\sigma)$ and $\beta_2(\sigma)$; regular altruists outperform Homo Hamiltonensis 1.0 at all $\epsilon \in [0, 1]$ for all $\sigma \in [0, 1]$ and all $\beta \in [\beta_1(\sigma), \beta_2(\sigma)]$.

Both types mix, and regular altruists still earn higher material payoffs than Homo Hamiltonensis 1.0 for $\beta > \beta_2(\sigma)$

For $\beta > \beta_2(\sigma)$, both a resident Homo Hamiltonensis 1.0 and a resident regular altruists would mix, at least at or close to $\epsilon = 0$. Also here, a resident Homo Hamiltonensis 1.0 can be invaded, while a resident regular altruist cannot (see also Figs. 5 and 6).

Summarizing, we can say that with the first way of allowing for mixing, we created non-empty sets of BNE for all β's, and we did get Homo Hamiltonensis 1.0 and regular altruists to behave differently, but as soon as that difference arises, regular altruists get *higher* (not lower) material payoffs than Homo Hamiltonensis 1.0.

4.2 Method II

In the example, we have introduced mixing by extending the definition of a BNE. At this step, we assumed expected utility, and thereby we created an extended version of Homo Hamiltonensis that we referred to as Homo Hamiltonensis 1.0. There is however also an alternative way to introduce mixing. We will refer to the Homo Hamiltonensis that we find with this alternative approach as Homo Hamiltonensis 2.0. The ways in which those are different will be discussed below.

An important observation, that we will make in Proposition 1 below, is that, while with the second approach, regular altruists stop outperforming Homo Hamiltonensis, also almost all of the difference between Homo Hamiltonensis and regular altruists goes away. More precisely, every equilibrium between two Homo Hamiltonensis 2.0 is also an equilibrium between regular altruists. Before we get to that result, we describe the intuition why Homo Hamiltonensis would be able to resist invasions without mixing, and how we can extend that intuition to mixing. That helps if we want to not only see how Homo Hamiltonensis 1.0 and 2.0 are different, but also why the second one would be able to resist invasions, and the first one would not, in case they lead to different behaviours.

4.3 The intuition behind stabiliy when all equilibria are pure

If we have a (fitness) game $\langle X, \pi \rangle$ in which X is a set of pure strategies, and for which there is exactly one pure BNE between every pair of types, the intuition for why Homo Hamiltonensis is evolutionarily stable, is beautifully uncomplicated. This intuition is described in Alger and Weibull (2013), and because this also plays an important role here, we will repeat it in a perhaps slightly more elaborate way.

Homo Hamiltonensis chooses an x that maximizes $(1-\sigma)\cdot\pi(x,y)+\sigma\cdot\pi(x,x)$. Being the resident, it only meets copies of itself at $\epsilon = 0$. Therefore, if x^* denotes what Homo Hamiltonensis plays in equilibrium at $\epsilon = 0$, this x^* is the x that maximizes $(1-\sigma)\cdot\pi(x,x^*)+\sigma\cdot\pi(x,x)$. This implies that Homo Hamiltonensis chooses an equilibrium strategy x^* against which the strategy that maximizes the material payoff of the mutant, would be to also play x^*. If we moreover assume that the best response is unique, as Alger and Weibull (2013) do in their central result, then every type that plays a strategy that is not x^* will have a material payoff, or invasion fitness, that is lower than the material payoff, or fitness, of Homo Hamiltonensis at $\epsilon = 0$. By choosing a strategy that already maximizes invasion fitness, Homo Hamiltonensis therefore preempts possible invaders.

Of course the material payoffs at $\epsilon = 0$ should also be informative about the payoffs for mutant shares ϵ close to 0. If we assume, for simplicity, that a resident Homo Moralis and mutant type τ, at sufficiently low shares of the mutant τ, always play a unique, pure, and symmetric BNE, that changes continuously as a function of ϵ, then there will be some $\bar{\epsilon}$ such that, if the share of mutants ϵ is below $\bar{\epsilon}$, the fitness of Homo Hamiltonensis is larger than that of the mutant, and the mutant will be pushed out. Making sure that payoffs change continuously, and also accommodating the possibility of multiple equilibria, complicates the formalizing of this intuition, but that is what Alger and Weibull (2013) do in their central result.[6]

A utility function that makes its carrier choose a strategy x^* that does *not* maximize $(1-\sigma)\cdot\pi(x,x^*)+\sigma\cdot\pi(x,x)$ by the same logic *is* vulnerable to invasion; if there is a strategy y^* for which $(1-\sigma)\cdot\pi(y^*,x^*)+\sigma\cdot\pi(y^*,y^*)$ is higher than $\pi(x^*,x^*)$, then a mutant strategy that would play this strategy y^* would be able to invade.

4.4 Expanding the intuition for when mixed equilibria are allowed for

For the more general setting, we would also consider the possibility that mutant preferences play mixed strategies, and that it would take a mixed equilibrium strategy to preempt invasions. In order to do that, we define Homo Hamiltonensis 2.0 as a preference that, when faced with an

[6] Accommodating multiple equilibria by requiring things to be true for all BNE is actually one of the root causes of the problems that arise if the set of BNE is empty.

individual that plays distribution ν, would choose a distribution μ that maximizes

$$(1-\sigma)\int\int \pi(x,y)d\mu(x)d\nu(y)+\sigma\int\int \pi(x,y)d\mu(x)d\mu(y) \tag{3}$$

As a resident, such a Homo Hamiltonensis 2.0 would, at $\epsilon = 0$, play a distribution μ^* that maximizes

$$(1-\sigma)\int\int \pi(x,y)d\mu(x)d\mu^*(y)+\sigma\int\int \pi(x,y)d\mu(x)d\mu(y) \tag{4}$$

This makes Homo Hamiltonensis 2.0 play the distribution that, against itself, maximizes invasion fitness.[7]

Homo Hamiltonensis 1.0, on the other hand, when facing a distribution ν, would choose x so as to maximize

$$(1-\sigma)\int \pi(x,y)d\nu(y)+\sigma\pi(x,x)$$

If it has multiple best responses, it can randomize over them, and since the value is the same for all best responses, any distribution over them would also be a μ that maximizes

$$(1-\sigma)\int\int \pi(x,y)d\mu(x)d\nu(y)+\sigma\int \pi(x,x)d\mu(x) \tag{5}$$

Randomizing in equilibrium can, obviously, be needed in order to construct a strategy that is a best response to itself, as it was in our example.

The difference between equations (3) and (5) is in the second term. The second term in (5) only considers symmetric strategy profiles (x,x), and weighs them with the probability that μ puts on different values for x. This way it reflects the hypothetical payoff one would get if both would choose the same *strategy*. The second term in (3) on the other hand weighs the payoffs for all possible strategy profiles with the probabilities with which they occur if both players randomize according to μ. This reflects the hypothetical expected payoff that one would get if both were to choose the same *distribution*. If $\sigma = 1$ – in which case Alger and Weibull (2013) would refer to Homo Hamiltonensis as Homo Kantiensis – then Homo Kantiensis 1.0 would choose the strategy that one would want to be chosen by everyone, and Homo Kantiensis 2.0 would choose the distribution that one would want to be chosen by everyone.

[7] Slightly more formally:

Observation 1. *If distribution μ^* maximizes (4), then the invasion fitness for any mutant type τ at any equilibrium (μ^*,ν^*) at $\epsilon = 0$ is less than or equal to the fitness of the resident at μ^*.*

Proof. The proof is straightforward; if there would be a type that played an equilibrium at $\epsilon = 0$ with a higher invasion fitness, then that contradicts μ^* being a maximum. □

The fact that Homo Hamiltonensis 2.0 as a resident by definition plays the distribution that would give the resident a fitness that matches the maximum invasion fitness implies that, if the second terms in (3) and (5) differ, then invasion fitness is not maximized at the equilibrium with a resident Homo Hamiltonensis 1.0, which therefore can be invaded.

There are of course settings in which the equilibrium behaviours of Homo Hamiltonensis 1.0 and 2.0 coincide. When restricted to a set of pure strategies, this difference obviously disappears, as it takes us back to the situation without mixing. There are moreover many combinations of games and distributions ν for which the distribution μ that maximizes (3) also maximizes (5) and vice versa.

4.5 Getting Homo Hamiltonensis 2.0 directly out of the model

We started with an example in which X is a set of pure strategies. We then expanded the Homo Hamiltonensis from that setting to also include lotteries. This we did in two different ways, one leading to Homo Hamiltonensis 1.0, and one to Homo Hamiltonensis 2.0. There is also a way in which one can get to Homo Hamiltonensis 2.0 more directly. Instead of defining the strategy set X as a set of pure strategies, one can also define X to be the set of probability distributions over that set of pure strategies. In our example, that would be if we choose X to be the set of distribution functions over $[1, \infty)$, instead of just $[1, \infty)$. This would then have to be combined with a restriction on π, which, with the new choice for X, is now defined over combinations of distribution functions. This restriction would have to be that it satisfies "expected material payoffs". If we momentarily stick to the notation in which x and y are pure strategies, and μ and ν are distributions over pure strategies, then $\pi(\mu, \nu) = \int \int \pi(\mathbf{1}_x, \mathbf{1}_y) d\mu(x) d\mu^*(y)$, where $\mathbf{1}_x$ and $\mathbf{1}_y$ are degenerate distributions on x and y, and $\pi(\mathbf{1}_x, \mathbf{1}_y)$ would have to be what $\pi(x, y)$ is for the setting with pure strategies only. We can then redefine the (fitness) game $\langle X, \pi \rangle$ so that X is now the set of distributions, and π is defined as we just did, provided that this new $\langle X, \pi \rangle$ still satisfies all the relevant requirements, such as X being a Hausdorff space.[8]

4.6 Any equilibrium between Homo Hamiltonensisses 2.0 is also an equilibrium between regular altruists

While Homo Hamiltonensis 2.0 cannot be invaded – unlike Homo Hamiltonensis 1.0 – this way of allowing for mixing does makes the "stark contrast" between altruists and Homo Hamiltonensis

[8] For consistency with the notation in Alger and Weibull (2013), one could then choose to rename distributions μ and ν, and call those x and y, and find new names for the pure strategies, that we referred to as x and y, because that is what they were in the case when X was just the set of pure strategies. Here we want to be able to go back and forth between these two, so we will continue to use μ and ν for distributions, and x and y for pure strategies.

go away. As we will see below, any equilibrium μ^* for a resident Homo Hamiltonensis 2.0 is also an equilibrium for a resident regular altruist with $\alpha = \sigma$. In order to see why, we can first consider equilibria in which such an equilibrium distribution μ has a density $f : X \to \mathbb{R}_0^+$.

For Homo Hamiltonensis 2.0, a necessary condition for distribution μ with density f to be optimal, when facing an individual that plays distribution ν with density $g : X \to \mathbb{R}_0^+$, would be that for all x with a positive density, the following would have to hold

$$(1-\kappa)\int \pi_1(x,y)g(y)dy + \kappa\left\{\int \pi_1(x,y)f(y)dy + \int \pi_2(y,x)f(y)dy\right\} = 0$$

where $\pi_1(x,y)$ is the derivative of π to its first, and $\pi_2(x,y)$ is the derivative of π to its second argument.

If this would not hold for a certain x at which $f(x) > 0$, then one can increase the value of (3), either by moving some probability mass from x to the left, if this expression is less than zero at x, or by moving probability mass to the right, if it is larger than zero. In equilibrium, this would then have to hold at $\nu = \mu$, or $g = f$, in which case we can rewrite this as

$$\int \pi_1(x,y)f(y)dy + \kappa\int \pi_2(y,x)f(y)dy = 0,$$

which would have to hold at all x for which $f(x) > 0$. That makes this expression constant 0, so this also implies

$$\frac{d}{dx}\left[\int \pi_1(x,y)f(y)dy + \kappa\int \pi_2(y,x)f(y)dy\right] = 0.$$

For a regular altruist with altruism parameter α, the necessary condition for optimality, for similar reasons, is

$$\int \pi_1(x,y)f(y)dy + \alpha\int \pi_2(y,x)f(y)dy = 0,$$

and for $\kappa = \alpha = \sigma$, those conditions are the same. For altruists, however, this moreover implies that all pure strategies present in the mix have the same utility, and, if none of the strategies not in the mix have a higher utility, that is also sufficient for it to be a Nash equilibrium. For the generalized Homo Moralis, this is not the case. For every pure strategy in the mix, the marginal contribution to the goal function may be the same, but that does not exclude the possibility that second order effects imply that this is not a maximum.

This also holds more generally.

Proposition 1. *If μ^* is an equilibrium for a resident Homo Hamiltonensis 2.0, then it is also an equilibrium for a resident regular altruist with $\alpha = \sigma$*

Proof. For a given μ, one can think of the marginal contribution of pure strategy x to the utility of a generalized Homo Hamiltonensis as

$$\begin{aligned} u_x &= (1-\sigma)\int \pi(x,y)d\mu(y) + \sigma\left(\int \pi(x,y)\mu(y) + \int \pi(y,x)d\mu(y)\right) \\ &= \int \pi(x,y)d\mu(y) + \sigma\int \pi(y,x)d\mu(y) \end{aligned}$$

A necessary condition for the utility function of a generalized Homo Hamiltonensis to be maximized, is that there should not be a pure strategy y and a set S with $\mu(S) > 0$ for which $u_y > \frac{\int_S u_x d\mu(x)}{\mu(S)}$; if there would be such a combination, then one could increase the utility by shifting mass away from S and to y. This implies that the marginal contribution is the same for almost all strategies in the support of μ, and that this is at least as high as the marginal contribution for all strategies not in the support of μ. This is also a necessary and sufficient condition for μ to be a symmetric Nash equilibrium between regular altruists with altruism parameter $\alpha = \sigma$. □

Proposition 1 only provides a one-way implication; all equilibria between Homo Hamiltonensis are also equilibria between regular altruists. The converse is not true. This leaves some space for examples in which a population of regular altruists that is playing a specific equilibrium can be invaded. If we consider a setting with 2 pure strategies, and allow for mixing over those, we can think of a coordination game.[9]

$$\begin{bmatrix} 1 & 0 \\ 0 & c \end{bmatrix}$$

with $c > 1$. We can formulate this so that it fits the setup of Alger and Weibull (2013) by letting x be the probability of playing the first strategy, and $1-x$ be the probability of playing the second strategy, which means that $\pi(x,y) = xy + c(1-x)(1-y)$. For regular altruists, there are two equilibria, independent of σ; $x = 1$ and $x = 0$. For Homo Hamiltonensis,[10] the latter is an equilibrium for all $\sigma \in [0,1]$ too, but $x = 1$ is an equilibrium only for $\sigma \leq \frac{1}{c}$. For $\sigma > \frac{1}{c}$, $x = 1$ is not an equilibrium for Homo Hamiltonensis, and, moreover, the equilibrium at $x = 1$ can be invaded by any strategy that plays $x = 0$, which includes Homo Hamiltonensis. If Homo

[9] We would like to thank an anonymous reviewer for this nice example.

[10] Since the set of pure strategies is finite, the version without mixing does not fit the setup of the model in Alger and Weibull (2013). If we do allow for mixing, we therefore refrain from calling the Homo Hamiltonensis with mixing Homo Hamiltonensis 2.0, as we did for a setting in which the game without mixing already fits the requirements of their central result. This example nonetheless illustrates the possibility that the set of equilibria for Homo Hamiltonenis with mixing is a strict subset of the equilibria for regular altruists. There are also examples where the game with pure strategies only already fits the requirements of the central result in Alger and Weibull (2013) – for instance the maximum effort game we get if we take the limit of $\beta \to \infty$ in our central example. We would be happy to provide such an example, but that would require more algebra.

Hamiltonensis would replace the regular altruists, it would however end up playing the $x = 0$ equilibrium, which is also an equilibrium between regular altruists.

In many cases, moreover, the equilibria for Homo Hamiltonensis 2.0 and regular altruists just coincide, as they for instance do in our example. With this approach to mixing, the blue lines in Fig. 3 come to represent both the equilibrium behaviour of a resident regular altruists, and the equilibrium behaviour of a resident Homo Hamiltonensis 2.0. More generally, Proposition 1 implies that if both have unique equilibria, they must coincide. Given that the remaining difference between Homo Hamiltonensis 2.0 and regular altruists only shows up when the former has multiple equilibria, one could consider thinking of evolutionary stability, not of preferences, but of combinations of a preference and an equilibrium. In that case, one could say that Homo Hamiltonensis 2.0 is evolutionarily stable at all of its equilibria. Regular altruists may have a larger set of equilibria, and equilibria for regular altruists that are not also equilibria for Homo Hamiltonensis 2.0 are not evolutionarily stable.

4.7 Overview

We started the example with a choice for $\langle X, \pi \rangle$ that satisfies the requirements for Alger and Weibull (2013)'s central result to apply. Their Theorem 1 implies that Homo Hamiltonensis – in this case having preferences over pure strategy profiles – satisfies their definition of evolutionary stability for our example. However, it does so by bypassing the comparison of material payoffs that is the basis of the indirect evolutionary approach that the authors claim to follow. There are some values of β for which the behaviour of Homo Hamiltonensis and regular altruists is well-defined, and identical, but for other β's, the behaviour of a mutant regular altruist is not defined, and for yet other β's neither the behavior of a mutant regular altruist, nor the behaviour of a resident Homo Hamiltonensis is defined. Their central result, applied to this example, correctly implies that Homo Hamiltonenis satisfies their definition of evolutionary stability for all β, but it is hard to maintain that this definition reflects the indirect evolutionary approach if the material payoffs that this method is meant to compare are not defined.

If we try to solve this problem by allowing for mixing with Method I, then the equilibrium behaviour of a resident Homo Hamiltonenis 1.0 and the equilibrium behaviour of a mutant regular altruist are defined for our example. Their behaviours are moreover proper different for $\beta > \beta_1$, but regular altruists now get *higher* instead of lower material payoffs.

If we try to solve this problem by allowing for mixing with Method II instead, then we find that Homo Hamiltonensis 2.0 does get material payoffs that mutants can at best match, but we also find that the difference between Homo Hamiltonensis 2.0 and regular altruists all but disappears. Proposition 1 shows that this is not something peculiar to our example, and

that for any game, any equilibrium between two Homo Hamiltonensis 2.0 is also an equilibrium between two regular altruists. In many cases (for instance if both have unique equilibria), their equilibrium behaviour as residents (and therefore also if one is the resident and the other a mutant) moreover simply coincides. This helps understand why, with mixing, regular altruists are evolutionarily stable for all $\beta > 0$ in our example. It also helps understand that if Homo Hamiltonensis 1.0 does something that regular altruists would not do, then Homo Hamiltonensis 1.0 can be invaded.

Finally, there are two details that might be worth mentioning. The first is that also if we allow for mixed strategies, as we did in Definition 4, then that is not a guarantee that BNE's will always exist (see Appendix 7.4). That should not come as a surprise, as Homo Hamiltonensis picks the (mixed) strategy with the highest invasion fitness, and since the existence of an ESS in strategy evolution is not guaranteed, there will be cases in which Homo Hamiltonensis will always be able to find a mutant strategy that can invade.

The second thing to notice is that in their paper, Alger and Weibull (2013) include an example in which strategies x and y themselves are interpreted as mixed strategies over a finite set of pure strategies. That fits their setup in the same way that the example at the end of Section 4.6 does. Such a setting does however come with restrictions on the set of payoff functions $\pi(x, y)$ – they would have to be bilinear, which is the same as being consistent with expected material payoffs if the set of pure strategies is finite. Therefore, if we choose a payoff function that is not bilinear – which is also allowed for – it cannot have this interpretation. The setup in Alger and Weibull (2013) therefore is open to the possibility that x represents a mixed strategy (as we have also seen in Section 4.5), but it is specifically also open to the possibility that x is something else, given the lack of further restrictions that would have to be imposed on π if x could *only* be interpreted as a distribution over pure strategies.

5 Coordination on asymmetric equilibria

The theory in Alger and Weibull (2013) does allow for some degree of coordination. When there are multiple symmetric pure BNE's, they assume that the population as a whole coordinates on one of them. What the current theory does not allow for, is coordination on asymmetric equilibria. Allowing for asymmetric equilibria in the set of BNE would however be an alternative way to reduce the number of instances for which sets of BNE are empty. Before we do exactly that, we go back to the setting without asymmetric equilibria, and without mixing. We consider the same game as in Section 3, but with a slightly larger strategy set; $X = [0, \infty)$ instead of $X = [1, \infty)$. The reason for the different strategy sets is just that $X = [1, \infty)$ gives simpler

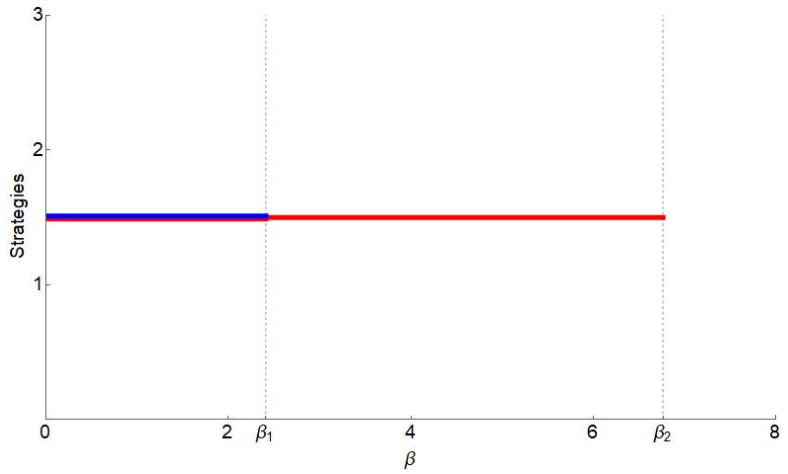

Figure 7: Behavior of a resident Homo Hamiltonensis (red) or a resident regular altruist (blue) in a Bayesian Nash equilibrium according to the definition in Alger and Weibull (2013) applied to our example, with $X = [0, \infty)$ and $\sigma = \frac{1}{2}$. At $\epsilon = 0$, the equilibrium behaviour does not depend on what type the mutant is.

equilibria with mixing, while $X = [0, \infty)$ gives more elegant equilibria if we allow for asymmetry. The change in X makes a bit of a difference for the way β_1 depends on σ, but other than that, the starting point for both is a situation with pure, symmetric BNE, where the equilibrium behaviour of a resident Homo Hamiltonensis and a resident regular altruist either coincides, or it is not defined for regular altruists, or for neither (Fig. 7).

5.1 Method I

There are two ways in which one can allow for asymmetric equilibria. In this subsection, we will discuss the first way. This produces non-empty sets of BNE, and equilibrium behaviour in which Homo Hamiltonensis and regular altruists choose different strategies. Between the two, regular altruists get the higher material payoffs whenever their equilibrium behaviour differs. After this subsection, we will discuss a second way of including allowing for asymmetric equilibria, but we will do this more generally before applying it to this example. With this second approach, there are also non-empty sets of BNE, but now the stark contrast in equilibrium behaviour between Homo Hamiltonensis and regular altruists goes away again.

A definition of a BNE that allows for asymmetric equilibria

With the first approach to allowing for asymmetric equilibria, we simply change the definition of a BNE. We do this by replacing strategies x^* and y^* by vectors (x_1^*, x_2^*) and (y_1^*, y_2^*) that specify what to play in role 1 and role 2, respectively. This also implies that there is a randomization

device that determines who gets to play which role in every pair. Half of the players of any type will then end up in one role, half in the other.

Definition 5. *In any state* $s = (\theta, \tau, \epsilon) \in S$*, a strategy pair* $((x_1^*, x_2^*), (y_1^*, y_2^*)) \in X^4$ *is a (Bayesian) Nash Equilibrium (BNE) if*

$$x_1^* \in \arg\max_{x \in X} Pr[\theta|\theta, \epsilon] \cdot u_\theta(x, x_2^*) + Pr[\tau|\theta, \epsilon] \cdot u_\theta(x, y_2^*),$$
$$x_2^* \in \arg\max_{x \in X} Pr[\theta|\theta, \epsilon] \cdot u_\theta(x, x_1^*) + Pr[\tau|\theta, \epsilon] \cdot u_\theta(x, y_1^*),$$
$$y_1^* \in \arg\max_{y \in X} Pr[\theta|\tau, \epsilon] \cdot u_\tau(y, x_2^*) + Pr[\tau|\tau, \epsilon] \cdot u_\tau(y, y_2^*)$$
$$y_2^* \in \arg\max_{y \in X} Pr[\theta|\tau, \epsilon] \cdot u_\tau(y, x_1^*) + Pr[\tau|\tau, \epsilon] \cdot u_\tau(y, y_1^*).$$

Symmetric equilibria are special cases, where $x_1^* = x_2^*$ and $y_1^* = y_2^*$, so this properly generalizes the definition of a BNE to include asymmetric equilibria.

Changing from Definition 1 to 5 implies that we have extended the domain of the utility functions of a Homo Hamiltonensis and a regular altruist, by assuming that the utility function is a weighted sum of the utilities that a player gets in role 1 and role 2 – where it is natural (but inconsequential) to choose those weights to be $\frac{1}{2}$. As with mixing, for a regular altruist, that is more or less the only reasonable option, but for Homo Hamiltonensis, as we will see below, there are also other ways to extend the domain of the utility function from a setting with symmetric equilibria only. We will use the same relatively short way of referring to the different versions of Homo Hamiltonensis as we did with mixing; the extended version of Homo Hamiltonensis implied by Definition 5 we will call Homo Hamiltonensis 1.0, and the alternative way of extending, that we will discuss later, we will refer to as Homo Hamiltonensis 2.0. Those are the same names as the extensions with mixing, but they are locally defined, in the sense that in Section 4 these names referred to different ways of allowing for mixed strategy profiles, and in this section they refer to different ways of allowing for double strategy profiles, that prescribe strategies played in both roles.

Equilibrium properties

If we do indeed allow for asymmetric equilibria in this way, then for $\beta > \beta_1(\sigma)$, a resident regular altruist in a BNE will choose different strategies for the different roles ($x_1^* \neq x_2^*$). Whether a BNE exists at $\epsilon = 0$ of course still depends on the mutant, but for any mutant that makes a BNE exist, the resident behavior is the same. This is depicted in Fig 8. In the interval $[\beta_{1'}, \beta_1(\sigma)]$,

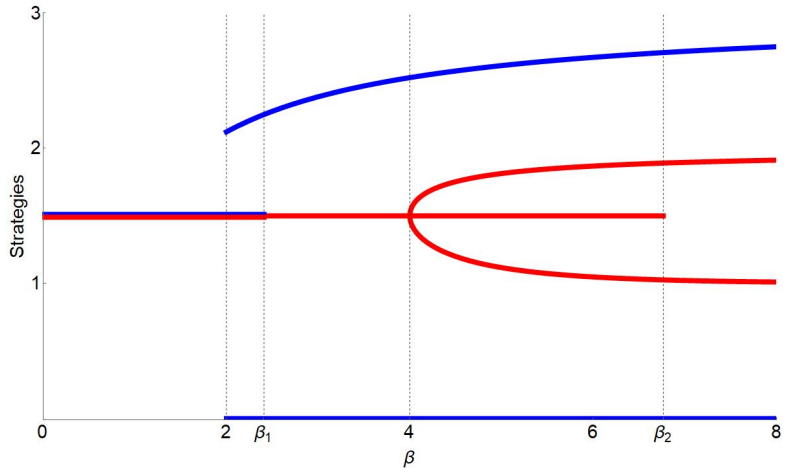

Figure 8: Equilibrium behavior of a resident Homo Hamiltonensis 1.0 (red) and a resident regular altruist (blue), both in Bayesian Nash Equilibria according to the definition in Alger and Weibull (2013), but extended to allow for asymmetric equilibria. At $\epsilon = 0$, the equilibrium behaviour does not depend on what type the mutant is. Between $\beta_{1'}$, which does not depend on σ, and equals 2, and β_1, there are two equilibria for a resident regular altruist; a symmetric one and an asymmetric one. Between $\beta_{2'}$, which does depend on σ, and equals 4 for $\sigma = \frac{1}{2}$, and β_2, this is true for a resident Homo Hamiltonensis.

moreover, there is both a symmetric and an asymmetric option for the equilibrium behavior of a resident regular altruist. The threshold $\beta_{1'}$ does not depend on σ, and equals 2.

All of this also applies to a resident Homo Hamiltonensis 1.0, but then with $\beta_{2'}(\sigma)$ and $\beta_2(\sigma)$ instead of $\beta_{1'}$ and $\beta_1(\sigma)$; there is an interval, $[\beta_{2'}(\sigma), \beta_2(\sigma)]$, in which both options are there, while beyond $\beta_2(\sigma)$, a resident Homo Hamiltonensis can only play asymmetrically. The threshold $\beta_{2'(\sigma)}$ does depend on σ, and equals 4 for $\sigma = \frac{1}{2}$.

The equilibrium behaviour and material payoffs of a resident Homo Hamiltonensis 1.0 is well-defined for all β, even though it is not always uniquely determined. The same is true for a resident regular altruist, and on top of this, they behave properly different for $\beta > \beta_1(\sigma)$. With the difference between them, comes the fact that a resident Homo Hamiltonensis 1.0 can be invaded, as soon as $\beta > \beta_1(\sigma)$, while regular altruists cannot. As a matter of fact, a resident Homo Hamiltonensis 1.0 can be invaded by a mutant regular altruist for $\beta > \beta_1(\sigma)$. Fig 9 depicts equilibrium strategies in a BNE between a resident Homo Hamiltonensis 1.0 and a mutant regular altruist at $\epsilon = 0$, and Fig. 10 shows that a mutant regular altruist always has *higher* material payoffs than a resident Homo Hamiltonensis 1.0 at invasion.

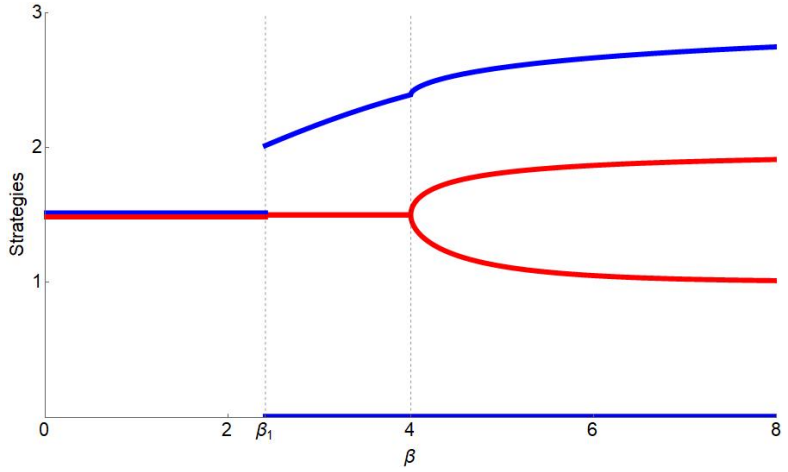

Figure 9: Equilibrium behaviour between a resident Homo Hamiltonensis 1.0 (red) and a mutant regular altruist (blue) at $\epsilon = 0$. In the interval in which there are two possibilities for a resident Homo Hamiltonensis 1.0 (a symmetric and an asymmetric one) we chose the asymmetric one, which brings higher material payoffs for the resident. Changing to the symmetric option for β between $\beta_{2'}(\sigma)$ and $\beta_2(\sigma)$ would of course also come with a change in behavior for the mutant regular altruist in the same interval, but in either instance, the mutant regular altruist outperforms the resident Homo Hamiltonensis 1.0 (see also Fig. 10).

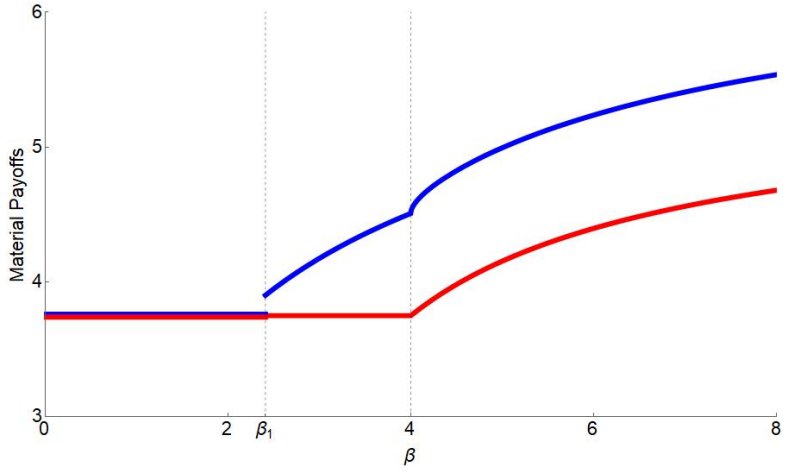

Figure 10: Material payoffs for the BNE, depicted in Fig. 9 between a resident Homo Hamiltonensis 1.0 (red) and a mutant regular altruist (blue) at $\epsilon = 0$.

5.2 Method II

Here, we define Homo Hamiltonensis 2.0 as a preference that, when faced with an individual that plays $y = (y_1, y_2)$, would choose a vector $x = (x_1, x_2)$, that maximizes

$$(1-\sigma)\frac{1}{2}\left(\pi(x_1, y_2) + \pi(x_2, y_1)\right) + \sigma\frac{1}{2}\left(\pi(x_1, x_2) + \pi(x_2, x_1)\right) \tag{6}$$

As a resident, such a Homo Hamiltonensis 2.0 would, at $\epsilon = 0$, play a vector $x^* = (x_1^*, x_2^*)$ that maximizes

$$(1-\sigma)\frac{1}{2}\left(\pi(x_1, x_2^*) + \pi(x_2, x_1^*)\right) + \sigma\frac{1}{2}\left(\pi(x_1, x_2) + \pi(x_2, x_1)\right) \tag{7}$$

This makes Homo Hamiltonensis 2.0 play the (possibly asymmetric) equilibrium that, against itself, maximizes invasion fitness.[11]

Homo Hamiltonensis 1.0, on the other hand, when facing a vector $y = (y_1, y_2)$, would choose x_1 so as to maximize

$$(1-\sigma)\pi(x_1, y_2) + \sigma\pi(x_1, x_1) \tag{8}$$

and x_2 so as to maximize

$$(1-\sigma)\pi(x_2, y_1) + \sigma\pi(x_2, x_2) \tag{9}$$

As a resident, such a Homo Hamiltonensis 1.0 would, at $\epsilon = 0$, choose an x_1^* that maximizes

$$(1-\sigma)\pi(x_1, x_2^*) + \sigma\pi(x_1, x_1) \tag{10}$$

and an x_2^* that maximizes

$$(1-\sigma)\pi(x_2, x_1^*) + \sigma\pi(x_2, x_2) \tag{11}$$

A symmetric equilibrium for a resident Homo Hamiltonensis 1.0 is also a symmetric equilibrium for a resident Homo Hamiltonensis 2.0, and vice versa. Asymmetric equilibria on the

[11] Slightly more formally:

Observation 2. *If $x^* = (x_1^*, x_2^*)$ maximizes (7), then the invasion fitness for any mutant type τ at any equilibrium (x^*, y^*) at $\epsilon = 0$ is less than or equal to the fitness of the resident at x^*.*

Proof. The proof is straightforward; if there would be a type that played an equilibrium at $\epsilon = 0$ with a higher invasion fitness, then that contradicts x^* being a maximum. □

other hand typically differ between the two. Also, from the utility functions, it is clear that for Homo Hamiltonenis 1.0, what matters are the hypothetical payoffs $\pi(x_1, x_1)$ and $\pi(x_2, x_2)$, and for Homo Hamiltonenis 2.0, what matters are the hypothetical payoffs $\pi(x_1, x_2)$ and $\pi(x_2, x_1)$. The fact that Homo Hamiltonensis 2.0 as a resident by definition plays the distribution that would give the resident a fitness that matches the maximum invasion fitness implies that, if the equilibrium behaviour of Homo Hamiltonensis 1.0 differs from that of Homo Hamiltonensis 2.0, then invasion fitness is not maximized at the equilibrium with a resident Homo Hamiltonensis 1.0, which therefore can be invaded.

5.3 Getting Homo Hamiltonensis 2.0 directly out of the model

There is also a way in which one can get to Homo Hamiltonensis 2.0 more directly. Instead of defining the strategy set X as a set of pure strategies, one can also define X as the product set of this set of pure strategies and the same set of pure strategies. In our example, that would be if we choose X to be $[0, \infty) \times [0, \infty)$, instead of just $[0, \infty)$. The new π would then also have to be defined over combinations of two elements of X, or, in other words, combinations $((x_1, x_2)(y_1, y_2))$ of four elements of $[0, \infty)$. The natural choice for a payoff function for our example would be $\pi((x_1, x_2)(y_1, y_2)) = \frac{1}{2}\left(a(x_1^\beta + y_2^\beta)^{\frac{1}{\beta}} - x_1^2 + a(x_2^\beta + y_1^\beta)^{\frac{1}{\beta}} - x_2^2\right)$. More in general, one would have to choose the payoff function π allowing for coordination on asymmetric equilibria so that it puts a positive weight on what the payoffs for strategy profiles (x_1, y_2) and (x_2, y_1) would be without coordination on asymmetric equilibria. A natural choice is for those weights to be equal. The restrictions this imposes on π are that $\pi((x_1, x_2), (y_2, y_1)) = \frac{1}{2}(\pi((x_1, y_2), (x_1, y_2)) + \pi((x_2, y_1), (x_2, y_1)))$.

If we redefine the (fitness) game $\langle X, \pi \rangle$ like this, then this fits the setup of the model from Alger and Weibull (2013) — provided that this new $\langle X, \pi \rangle$ still satisfies all the relevant requirements, such as X being a Hausdorff space.

5.4 Any equilibrium between Homo Hamiltonensissis 2.0 is also an equilibrium between regular altruists

Here we show that any equilibrium $x^* = (x_1^*, x_2^*)$ for a resident Homo Hamiltonensis 2.0 is also an equilibrium for a resident regular altruist with $\alpha = \sigma$. In this case, this is relatively straightforward.

Proposition 2. *If $x^* = (x_1^*, x_2^*)$ is an equilibrium for a resident Homo Hamiltonensis 2.0, then it is also an equilibrium for a resident regular altruist with $\alpha = \sigma$.*

Proof. If $x^* = (x_1^*, x_2^*)$ is an equilibrium for a resident Homo Hamiltonensis 2.0, then it is a fixed point, where $(x_1, x_2) = (x_1^*, x_2^*)$ maximizes

$$(1-\sigma)\frac{1}{2}\left(\pi(x_1, x_2^*) + \pi(x_2, x_1^*)\right) + \sigma\frac{1}{2}\left(\pi(x_1, x_2) + \pi(x_2, x_1)\right)$$

Two necessary conditions for this to be true are that x_1 maximizes this expression at $x_1 = x_1^*$ if x_2 is fixed at x_2^*, and that x_2 maximizes this expression at $x_2 = x_2^*$ if x_1 is fixed at x_1^*. The first of those therefore requires that

$$(1-\sigma)\frac{1}{2}\left(\pi(x_1, x_2^*) + \pi(x_2^*, x_1^*)\right) + \sigma\frac{1}{2}\left(\pi(x_1, x_2^*) + \pi(x_2^*, x_1)\right)$$

is maximized at $x_1 = x_1^*$ which is the same as

$$\pi(x_1, x_2^*) + \sigma\pi(x_2^*, x_1)$$

being maximized at $x_1 = x_1^*$. In combination with its mirror image for fixing x_1 at $x_1 = x_1^*$, these also define what a fixed point is for a regular altruist. □

As Proposition 1 does with mixing, Proposition 2 only provides a one-way implication; all equilibria between Homo Hamiltonensis are also equilibria between regular altruists. The converse is not true; there can be equilibria between regular altruists that are not equilibria between Homo Hamiltonensis 2.0. That means that at such an equilibrium, these could be invaded. A Homo Hamiltonensis 2.0, however, will always settle at behaviour that is also equilibrium behaviour between regular altruists. Also, in many cases the equilibria just coincide. In our example, Homo Hamiltonensis 2.0 matches the equilibrium behaviour of regular altruists (the blue lines in Fig. 8) when there is a unique equilibrium, and between $\beta_{1'}$ and β_1, the asymmetric equilibrium for a regular altruist is also an equilibrium for Homo Hamiltonensis 2.0, but the symmetric equilibrium is not.

6 Discussion

Alger and Weibull (2013) set out to answer the question whether in a setting with assortment, we should expect altruistic preferences to evolve, or moral ones. Their central result suggests that this evolutionary competition is won by the latter. Under conditions specified in the paper, their central result states that Homo Hamiltonensis satisfies their definition of evolutionary stability, and preferences that make one behave differently are evolutionarily unstable. They also suggest that there can be a "stark contrast" between the behaviour of Homo Hamiltonensis and regular

altruists.

We find that this paints an incorrect picture. First of all, it is important to realize that Homo Hamiltonensis can satisfy their definition of evolutionary stability through a loophole, when the set of Bayesian Nash Equilibria for a resident Homo Hamiltonensis and any mutant is empty. In this case the equilibrium behaviour of a resident Homo Hamiltonensis is not defined, so any statement about its evolutionary stability is not really based on behaviour or material payoffs. Also when the equilibrium behaviour of a resident Homo Hamiltonensis is defined, but the behaviour of a mutant regular altruist is not, it is not justified to state that the first is evolutionarily stable against the second.

If we try to close those loopholes by allowing for mixing, or for coordination on asymmetric equilibria, then this can be done in two ways. The first creates a stark contrast, but then regular altruists win the evolutionary competition. The second makes Homo Hamiltonensis not invadable, but then the contrast between Homo Hamiltonensis and regular altruists all but disappears; our Propositions 1 and 2 imply that every equilibrium for a Homo Hamiltonensis 2.0 (the version for which it cannot be beaten by regular altruists) is also an equilibrium between regular altruists.

These propositions are only one-way implications, and it is in fact possible to find games for which there is an equilibrium between regular altruist that can be invaded. However, if Homo Hamiltonensis 2.0 would take over the population, then any equilibrium that it could play is also an equilibrium between regular altruists. At those equilibria, regular altruists moreover could not be invaded. Also, for many games, their equilibria just coincide; for instance for games in which both have unique equilibria, Propositions 1 and 2 imply that these must coincide.

The options in a setting with assortment and imperfect information, therefore, are that either there is a stark contrast, but then regular altruists beat Homo Hamiltonensis, or Homo Hamiltonensis cannot be beaten, but then almost all of the contrast between morality and altruism dissipates.

7 Appendices

7.1 First order conditions and thresholds

Homo Moralis The utility function of Homo Moralis is $u_\kappa = (1-\kappa)\cdot\pi(x,y)+\kappa\cdot\pi(x,x)$. If we assume that the material payoff function π is continuous and twice differentiable, then the first order condition for maximizing it with respect to x is

$$(1-\kappa)\cdot\pi_1(x,y)+\kappa\cdot(\pi_1(x,x)+\pi_2(x,x))=0$$

where π_1 is the derivative of π to its first argument, and π_2 is the derivative of π to its second argument. In a symmetric, pure equilibrium, where $x = y$, this can also be written as

$$\pi_1(x,x) + \kappa \cdot \pi_2(x,x) = 0$$

Regular altruist The utility function of a regular altruist is $u_\alpha = \pi(x,y) + \alpha \cdot \pi(y,x)$. The first order condition for maximizing this utility function with respect to x is

$$\pi_1(x,y) + \alpha \cdot \pi_2(y,x) = 0$$

In a symmetric pure equilibrium, where $x = y$, this can also be written as

$$\pi_1(x,x) + \alpha \cdot \pi_2(x,x) = 0$$

If $\alpha = \kappa = \sigma$, these first order conditions are the same.

7.2 Second order conditions

Homo Moralis For Homo Moralis, the second order condition requires that $(1-\kappa)\cdot\pi_{1,1}(x,y) + \kappa\cdot(\pi_{1,1}(y,x) + \pi_{1,2}(y,x) + \pi_{2,1}(y,x) + \pi_{2,2}(y,x))$ is less than 0 at $x = y$, or, in other words, that

$$\pi_{1,1}(x,x) + \kappa\cdot(2\pi_{1,2}(x,x) + \pi_{2,2}(x,x)) < 0$$

.

Regular altruist For regular altruists, the second order condition requires that $\pi_{1,1}(x,y) + \alpha\cdot\pi_{2,2}(y,x)$ is less than 0 at $x = y$, or, in other words,

$$\pi_{1,1}(x,x) + \alpha\cdot\pi_{2,2}(x,x) < 0$$

The second order conditions therefore are not the same, even if $\alpha = \kappa = \sigma$. It is moreover possible that there is an x for which the first and second order conditions are satisfied, while x is a local, but not a global maximum. This is the case for some β's in our example.

7.3 Our example

The first order condition for a Homo Moralis with morality parameter $\kappa = \sigma$, or a regular altruist with altruism parameter $\alpha = \sigma$, in combination with our example, $\pi(x,y) = a(x^\beta + y^\beta)^{\frac{1}{\beta}} - x^2$, gives:

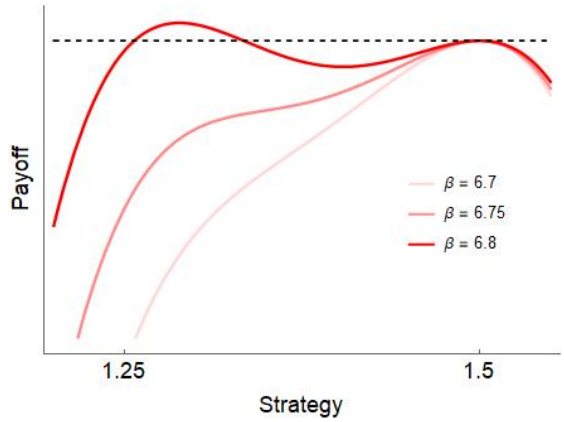

Figure 11: Utilities for Homo Hamiltonensis when $\sigma = \frac{1}{2}$ and $x^* = 1 + \sigma$, and for different β's. For $\beta = 6.7$ and $\beta = 6.75$, $x^* = 1.5$ is a fixed point, but for $\beta = 6.8$ it is not.

$$\left[a \cdot \beta x^{\beta-1} \frac{1}{\beta}(x^\beta + y^\beta)^{\frac{1}{\beta}-1} - 2x\right]_{y=x} + \sigma\left[a \cdot \beta y^{\beta-1} \frac{1}{\beta}(x^\beta + y^\beta)^{\frac{1}{\beta}-1}\right]_{y=x} = 0$$

$$a\,(1+\kappa)\left[x^{\beta-1}(2x^\beta)^{\frac{1}{\beta}-1}\right] - 2x = 0$$

$$a\,(1+\sigma)\,2^{\frac{1}{\beta}-1} = 2x$$

With $a = 2^{2-\frac{1}{\beta}}$, this makes

$$x = 1 + \sigma$$

That implies that for our example, the only candidate for a pure, symmetric equilibrium is $x = 1 + \sigma$.

One can show that the second order condition for a Homo Hamiltonensis holds if $\beta < \frac{3+\sigma}{1-\sigma}$. There is however a range of β's for which the second order condition still holds at $x = 1 + \sigma$, but for which, if we take $x^* = 1 + \sigma$, $x = 1 + \sigma$ is not a global maximum (see Fig. 11). That implies that there is no fixed point, and the β at which that starts happening is $\beta_2(\sigma)$. For $\sigma = \frac{1}{2}$, this threshold is $\beta_2\left(\frac{1}{2}\right) \approx 6.792$.

For a regular altruist, one can show that the second order condition holds if $\beta < 3$. Also here, there is a range of β's for which the second order condition holds at $x = 1+\sigma$, but for which, if we take $x^* = 1 + \sigma$, $x = 1 + \sigma$ is not a global maximum. Where that starts happening, depends on X and on σ (with one exception). For $X = [0, \infty)$, $\beta_1 = \frac{\ln 2}{\ln 4 - \ln 3}$, which happens to be independent of σ. For $X = [1, \infty)$, β_1 does depend on σ, and for $\sigma = \frac{1}{2}$ it is $\beta_1\left(\frac{1}{2}\right) \approx 2.72$.

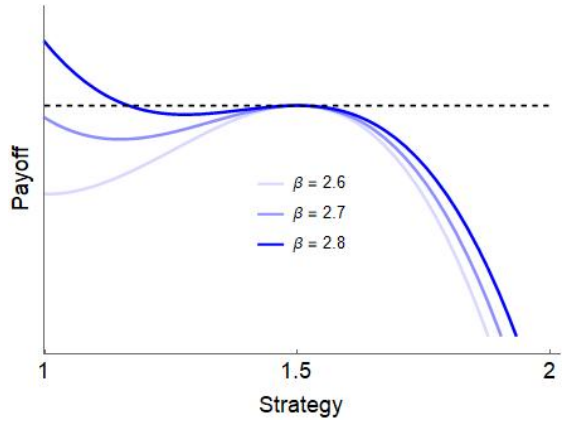

Figure 12: Utilities for a regular altruist when $X = [1, \infty)$, $\sigma = \frac{1}{2}$, and $x^* = 1 + \sigma$, and for different β's. For $\beta = 2.6$ and $\beta = 2.7$, $x^* = 1.5$ is a fixed point, but for $\beta = 2.8$ it is not.

If we allow for mixing, as we do in Section 4.1, then regular altruists and Homo Hamiltonensis will play pure strategies for values of β below $\beta_1(\sigma)$ and $\beta_2(\sigma)$, respectively, and mixed strategies for values of β above $\beta_1(\sigma)$ and $\beta_2(\sigma)$. If we allow for coordination on asymmetric equilibria, as we do in Section 5.1, then symmetric equilibria stop existing at the same thresholds. Asymmetric equilibria however start existing for lower values of β, and therefore there are intervals, with $\beta_1(\sigma)$ and $\beta_2(\sigma)$ as their respective upper bounds, in which both mixed and pure equilibria exist. For regular altruist, the lower bound of this interval is $\beta_{1'} = 2$. This bound is independent of X and σ. For Homo Hamiltonensis, the lower bound does depend on σ – but not on X – and for $\sigma = \frac{1}{2}$ it is $\beta_{2'}\left(\frac{1}{2}\right) = 4$.

7.4 A game with mixing, but without BNE at $\epsilon = 0$

Consider the Rock-Scissors-Paper game, parametrized as in Weibull (1997), here with $a \in (\sigma - 1, 0)$

$$\begin{bmatrix} 1 & 2+a & 0 \\ 0 & 1 & 2+a \\ 2+a & 0 & 1 \end{bmatrix}$$

This is not a game that fits the setup of the model in Alger and Weibull (2013) in the absence of mixing, but with mixing it nonetheless illustrates that mixing does not guarantee the existence of a BNE for a Homo Hamiltonensis. With strategies $x = [x_1, x_2, x_3]$ and $y = [y_1, y_2, y_3]$, we therefore consider the utility function of a type θ that is a Homo Hamiltonensis:

$$
\begin{aligned}
u_\theta(x,y) &= (1-\sigma)[x_1,x_2,x_3]\begin{bmatrix} 1 & 2+a & 0 \\ 0 & 1 & 2+a \\ 2+a & 0 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\
&\quad + \sigma[x_1,x_2,x_3]\begin{bmatrix} 1 & 2+a & 0 \\ 0 & 1 & 2+a \\ 2+a & 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \\
&= 1 + (1-\sigma)\left((1+a)(x_1y_2 + x_2y_3 + x_3y_1) - (x_1y_3 + x_2y_1 + x_3y_2)\right) \\
&\quad + \sigma a(x_1x_2 + x_2x_3 + x_3x_1)
\end{aligned}
$$

A fixed point x^* therefore requires that x^* maximizes

$$
\begin{aligned}
u_\theta(x,x^*) &= 1 + (1-\sigma)\left((1+a)(x_1x_2^* + x_2x_3^* + x_3x_1^*) - (x_1x_3^* + x_2x_1^* + x_3x_2^*)\right) \\
&\quad + \sigma a(x_1x_2 + x_2x_3 + x_3x_1)
\end{aligned}
$$

For a fixed point in the interior of the simplex, a necessary condition would be that

$$
\frac{du_\theta(x,x^*)}{dx_1}\Big|_{x=x^*} = \frac{du_\theta(x,x^*)}{dx_2}\Big|_{x=x^*} = \frac{du_\theta(x,x^*)}{dx_3}\Big|_{x=x^*}
$$

In other words, at $x = x^*$, the following should hold:

$$
\begin{aligned}
&(1-\sigma)\left[(1+a)x_2^* - x_3^*\right] + \kappa\left[a(x_2+x_3)\right] \\
&= (1-\sigma)\left[(1+a)x_3^* - x_1^*\right] + \kappa\left[a(x_1+x_3)\right] \\
&= (1-\sigma)\left[(1+a)x_1^* - x_2^*\right] + \kappa\left[a(x_1+x_2)\right]
\end{aligned}
$$

This implies that all three probabilities should be equal; $x = x^* = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$. However, utility is not maximized there; with $x^* = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$, we find that

$$
\begin{aligned}
u_\theta(x,x^*) &= 1 + (1-\sigma)\left((1+a)\frac{1}{3}(x_1+x_2+x_3) - \frac{1}{3}(x_1+x_2+x_3)\right) \\
&\quad + \sigma a(x_1x_2 + x_2x_3 + x_3x_1) \\
&= 1 + (1-\sigma)\cdot a\cdot\frac{1}{3} + \sigma a(x_1x_2 + x_2x_3 + x_3x_1)
\end{aligned}
$$

The first terms are independent of x, and for $x = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$ the last term is $\frac{1}{3}\cdot\kappa\cdot a < 0$, while for $x = [1,0,0]$, $x = [1,0,0]$, or $x = [1,0,0]$, the last term is 0, which makes the utility for any of the

three pure strategies higher than the utility at $x = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right]$.

If we take $x^* = [x_1^*, 1 - x_1^*, 0]$, then we find, again by filling in x^*, that

$$\begin{aligned} u_\theta(x, x^*) = 1 + (1-\sigma)\left((1+a)(x_1(1-x_1^*) + x_3 x_1^*) - (x_2 x_1^* + x_3(1-x_1^*))\right) \\ + \sigma a(x_1 x_2 + x_2 x_3 + x_3 x_1) \end{aligned}$$

For a fixed point on this edge of the simplex, a necessary condition would be that

$$\frac{du_\theta(x, x^*)}{dx_1}|_{x=x^*} = \frac{du_\theta(x, x^*)}{dx_2}|_{x=x^*} < \frac{du_\theta(x, x^*)}{dx_3}|_{x=x^*}$$

In other words, the following should hold at $x = x^*$:

$$(1-\sigma)\left[(1+a)(1-x_1^*)\right] + \sigma\left[a(x_2 + x_3)\right] = (1-\sigma)\left[-x_1^*\right] + \sigma\left[a(x_1 + x_3)\right]$$

At $x = x^* = [x_1^*, 1 - x_1^*, 0]$, this would be

$$\begin{aligned} (1-\sigma)\left[(1+a)(1-x_1^*)\right] + \sigma\left[a(1-x_1^*)\right] &= (1-\sigma)\left[-x_1^*\right] + \sigma\left[a(x_1^*)\right] \\ (1-\sigma)(1-x_1^*) + a(1-x_1^*) &= (1-\sigma)\left[-x_1^*\right] + \sigma\left[a(x_1^*)\right] \\ 1 - \sigma + a &= (1+\sigma)\left[a(x_1^*)\right] \\ \frac{1-\sigma+a}{(1+\sigma)a} &= x_1^* \end{aligned}$$

This would be negative for $a \in (\sigma - 1, 0)$.

None of the three pure strategies are BNE either. If we take for instance $x^* = [1, 0, 0]$, then

$$u_\theta(x, x^*) = 1 + (1-\sigma)\left((1+a)x_3 - x_2\right) + \sigma a(x_1 x_2 + x_2 x_3 + x_3 x_1)$$

For $x = [1, 0, 0]$, this is $u_\theta(x, x^*) = 1$, but for $x = [0, 0, 1]$ this is $u_\theta(x, x^*) = 1 + (1-\sigma)(1+a) > 1$.

References

Akdeniz, A., C. Graser, and M. van Veelen (2020): "Homo Moralis and regular altruists–preference evolution for when they disagree," *Tinbergen Institute Discussion Paper 2020-062/I.*

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

Dawkins, R. (1976): *The selfish gene*. Oxford University Press.

Hamilton, W. D. (1964): "The genetical evolution of social behaviour. I," *Journal of theoretical biology*, 7(1), 1–16.

Kay, T., L. Keller, and L. Lehmann (2020): "The evolution of altruism and the serial rediscovery of the role of relatedness," *Proceedings of the National Academy of Sciences*, 117(46), 28894–28898.

van Veelen, M. (2009): "Group selection, kin selection, altruism and cooperation: when inclusive fitness is right and when it can be wrong," *Journal of Theoretical Biology*, 259(3), 589–600.

——— (2011): "The replicator dynamics with n players and population structure," *Journal of Theoretical Biology*, 276(1), 78–85.

——— (2018): "Can Hamilton's rule be violated?," *eLife*, 7, e41901.

van Veelen, M., B. Allen, M. Hoffman, B. Simon, and C. Veller (2017): "Hamilton's rule," *Journal of theoretical biology*, 414, 176–230.

Weibull, J. W. (1997): *Evolutionary game theory*. MIT press.