TI 2020-062/I
Tinbergen Institute Discussion Paper

Homo Moralis and regular altruists – preference evolution for when they disagree

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

[1] University of Amsterdam

Homo Moralis and regular altruists – preference evolution for when they disagree

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

[1]University of Amsterdam, The Netherlands.
[2]Tinbergen Institute, The Netherlands.

September 17, 2020

Abstract

Alger and Weibull (2013) present a model for the evolution of preferences under incomplete information and assortative matching. Their main result is that Homo Moralis – who maximizes a convex combination of her narrow self-interest and "the right thing to do" – is evolutionarily stable, if it assigns a weight on the right thing to do that is equal to the assortment parameter. We give a counterexample against their central result, and a way to repair it. We also show that the result ceases to hold if we allow for mixed equilibria or coordination on asymmetric equilibria. Allowing for mixed equilibria, we show that if there is a stable preference, it will be behaviorally equivalent to a regular altruist that puts a positive weight on the payoff of the other that is equal to the assortment parameter. We also consider the cross-species empirical evidence.

Keywords Homo Moralis, altruism, preference evolution

1 Introduction

How human morality evolved is one of the bigger questions. In a recent paper, Alger and Weibull (2013) suggest that the key to the answer might lie in a combination of assortative matching and imperfect information about each other's preferences. In their model of preference evolution, they show that a type that they call Homo Hamiltonensis is evolutionarily stable against all types that are not behavioral alikes. Homo Hamiltonensis is a special case of what they call Homo Moralis. The utility function of a Homo Moralis is a convex combination of her own material

interests and a hypothetical payoff, which is the material payoff that one would get if both would play the strategy that one plays oneself. Maximizing the latter in symmetric games can also be described as "doing the right thing", or following Kant's categorical imperative (Kant, 1785). Homo Hamiltonensis is a Homo Moralis that puts a weight on "the right thing to do" that is equal to the assortment parameter. The name is a reference to William Hamilton, and *Hamilton's rule*, which is a famous result in theoretical biology, in which assortment, or relatedness, also plays a central role (Hamilton, 1964a,b).

The key ingredients of the model are assortment and imperfect information, but along the way a few restrictions are made. Two of those are that the central result only considers cases where all equilibria are pure, and, while coordination on symmetric equilibria is allowed for, coordination on asymmetric ones is not. These restrictions do not seem essential, but we will show that the result ceases to hold if they are lifted, and Homo Hamiltonensis can become unstable.

Before we get to these restrictions, we will first reproduce the model, which is done in Section 2, and discuss two important technical points, which we will do in Sections 3 and 4. In Section 3, we will give a counterexample against the result. That means that the result, in the form presented in Alger and Weibull (2013) is not correct. We also suggest a way to repair the result, which is to restrict the set of admissible mutants. In Section 4, we give an example of a behavioral alike of Homo Hamiltonensis that can invade. A behavioral alike is a type for which at frequency 0 there is at least one equilibrium at which her behaviour is indistinguishable from that of Homo Hamiltonensis. That suggests that behavioral alikes might be harmless, but the example shows that behaving the same at a mutant frequency of 0 does not imply behaving the same at other mutant frequencies, and we will see that some behavioral alikes therefore may be able to invade.

The lifting of the restriction to pure equilibria happens in Section 5. Allowing for mixed equilibria, we find that Homo Hamiltonensis is no longer evolutionarily stable against strategies that are not behavioral alikes. We show that there is a different kind of Homo Moralis that can be considered relatively stable, in the sense that there is no preference that has an invasion fitness that is larger than the fitness of this other Homo Moralis. The difference between the Homo Moralis from Alger and Weibull (2013) and this Homo Moralis is that for a mixed strategy, this one puts a positive weight on the expected payoff in case both players would choose that mix, while the original Homo Moralis puts a positive weight on the expected payoff if one of them would randomize according to that mix, and the other player would always match the pure strategy that results from the randomization. We also show that any equilibrium that this generalized Homo Moralis could play as a resident is also an equilibrium between regular

altruists. That explains why, if the original and the generalized Homo Moralis differ, also regular altruists are more stable than the original Homo Moralis.

In Section 6 we allow for coordination also on asymmetric equilibria, which also implies that Homo Hamiltonensis becomes unstable, and regular altruists do not.

In Section 7 we will look at the cross-species empirical evidence to see if it fits the idea that morality is the product of a combination of assortment and incomplete information. This turns out to be mixed at best. Humans, presumably the only species in which morality has evolved, are not special in their population structure, as far as genetic assortment is concerned. We are however relatively good at kin recognition, and culture allows for ways to generate assortment that other species may not have. Human morality on the other hand also extends to individuals with whom relatedness is 0, or, in other words, to interactions that are not assorted. Also humans are specializing in theory of mind, so they are typically better, not worse, than other animals at inferring and understanding each other's preferences.

2 The Model

Alger and Weibull (2013) present a model for the evolution of preferences under incomplete information and assortative matching. They consider a population where individuals are matched in pairs to engage in a symmetric interaction with a common strategy set X (see Alger and Weibull (2016) for an extension to n-player games. These individuals have preferences, which determine what they choose to do in the interaction. What they do in the interaction determines their evolutionary success, according to a 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 a resident, that is, $\lim_{\epsilon \downarrow 0} Pr[\theta|\theta, \epsilon] = 1$, while the mutant is matched with a resident with probability $\lim_{\epsilon \downarrow 0} Pr[\theta|\tau, \epsilon] = 1 - \sigma$, and with another mutant with probability $\lim_{\epsilon \downarrow 0} Pr[\tau|\tau, \epsilon] = \sigma$. 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 themselves with probability σ.[1] Since a random draw at mutant frequency 0 means being

[1] 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 $\epsilon > 0$.

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, and residents know what the preferences of the mutant are. 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 type, and where Θ is the set of types we are considering, which 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 $B^{NE}(\theta, \tau, \epsilon)$. 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 $\epsilon = 0$.

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

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 $\epsilon = 0$. This is reflected in the notation for the payoffs at $\epsilon = 0$ that we introduce here, and that is not in the original paper. 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 denote their payoffs as follows.

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

We will call $\Pi_\theta(x^*)$ the fitness of the resident θ for x^*, and we will 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 we will call $\Pi_\theta = \Pi_\theta(x^*)$ the fitness of the resident. This fitness is 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 we will 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.

The main result in Alger and Weibull (2013) is that, under some restrictions 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 is a special case of what they label a Homo Moralis, who 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 2. *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 Hamiltonensis plays against a copy of herself, 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.

2.1 The intuition behind the result

If Homo Hamiltonensis is evolutionarily stable against another type that is not behaviorally equivalent, then the intuition for why that is, is not complicated. It is described in Alger and Weibull (2013), and because this will also play an important role later on, in adapted form, we will repeat it here, 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, she only meets copies of herself 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 Hamiltonensis and mutant type τ, when facing a mix of them with sufficiently few mutants τ in it, 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) managed to do in their central result.

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.

2.2 Regular altruists are often behaviourally equivalent

Alger and Weibull (2013) also point out that for some combinations of a payoff function π and a strategy that an opponent could play, the strategy that Homo Hamiltonensis chooses is the same as the strategy that regular altruists, with an altruism parameter equal to the assortment parameter, would choose. They point to the first order conditions being the same, and because

their behavioural equivalence will also play a role later on, in a slightly generalized way, we will do the same.

The utility function of Homo Moralis is $u_\kappa = (1-\kappa)\cdot\pi(x,y) + \kappa\cdot\pi(x,x)$, and if we assume that the 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

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

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

The second order conditions are not the same. For Homo Hamiltonensis, 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 $\pi_{1,1}(x,x) + \kappa\cdot(2\pi_{1,2}(x,x) + \pi_{2,2}(x,x)) < 0$. 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, $\pi_{1,1}(x,x) + \alpha\cdot\pi_{2,2}(x,x) < 0$. These conditions are not the same, even if $\alpha = \kappa = \sigma$.

If x^* is the equilibrium behaviour of a resident Homo Hamiltonensis – which would imply that both first and second order conditions are satisfied at x^* – then it will also constitute equilibrium behaviour of a resident regular altruist with $\alpha = \sigma$ as long as $\pi_{1,2}(x^*,x^*)) < -\left(\frac{1}{\sigma}\pi_{1,1}(x^*,x^*) + \pi_{2,2}(x^*,x^*)\right)$, or, in other words, as long as the strategic complementarity is not so strong that the second order condition does not hold anymore. Also, Homo Hamiltonensis and regular altruists with $\alpha = \sigma$ then would behave indistinguishably in any mixture of them.

3 A counterexample against the main result

The first part of Theorem 1 in Alger and Weibull (2013) reads as follows:

"If $\beta_\sigma(x)$ is a singleton for all $x \in X_\sigma$, then homo hamiltonensis *is evolutionarily stable against all types $\tau \notin \Theta_\sigma$."*

Here, X_σ refers to the set of fixed points of Homo Hamiltonensis' best response correspondence $\beta_\sigma(x)$, and Θ_σ refers to the set of "behavioral alikes" of Homo Hamiltonensis, which is defined as follows:

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

In other words, a preference is a behavioral alike as soon as there is a BNE at $\epsilon = 0$ at which the mutant and the resident play the same (pure) strategy.

3.1 Example 1

Consider the game

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

For $\beta \to \infty$, this game converges to what one could call the "maximum effort game"; $\pi(x,y) = a\max\{x,y\} - x^2$. This game also features in Sections 5 and 6. To make the solutions easy to read, we will choose $a = 2^{2-\frac{1}{\beta}}$.

The first order condition for a symmetric, pure equilibrium would imply that a resident Homo Moralis with morality parameter $\kappa = \sigma$ would choose $x^* = 1+\sigma$ (the derivation is in Appendix A). The second order condition for a resident Homo Moralis requires that $\beta < \frac{3+\sigma}{1-\sigma}$ (this derivation is also in Appendix A). If we choose $\kappa = \sigma = \frac{1}{2}$ and $\beta = 4$, then both are satisfied if Homo Hamiltonensis plays $x^* = 1+\sigma = \frac{3}{2}$. The set of fixed points X_σ of Homo Hamiltonensis' best response correspondence is the singleton set $\{\frac{3}{2}\}$, and Homo Hamiltonesis' best response $\beta_\sigma(x)$ is a singleton for $x = \frac{3}{2}$, so the condition stated in the theorem is satisfied. Yet, if we take as a mutant a preference that in one of the equilibria randomizes between a value $x_0 < \frac{3}{2}$ and a value $x_1 > \frac{3}{2}$, then it can, in such an equilibrium, have a fitness that is higher than the fitness of the resident Homo Hamiltonensis. Consider a preference that is represented by the utility function $u_\tau(x,y) = -(x-1.45)^2$ if $x \leq \frac{3}{2}$ and $u_\tau(x,y) = -(x-1.55)^2$ if $x > \frac{3}{2}$. This preference is independent of the action of the other, and is indifferent between any way it randomizes between $x_0 = 1.45$ and $x_1 = 1.55$. Let p be the probability with which the mutant plays x_0 and $1-p$ the

probability with which the mutant plays x_1.

The fitness of the resident Homo Hamiltonensis at $\epsilon = 0$ is

$$\Pi_\sigma = \pi\left(\frac{3}{2}, \frac{3}{2}\right) = 2^{2-\frac{1}{4}} \cdot \left(2 \cdot \left(\frac{3}{2}\right)^4\right)^{\frac{1}{4}} - \frac{9}{4} = \frac{15}{4}$$

The fitness of the mutant at the equilibrium described, with assortment parameter σ, is given by

$$\begin{aligned} &p \cdot \left(\sigma \cdot a \cdot \left(x_0^4 + \left(\frac{3}{2}\right)^4\right)^{\frac{1}{4}} + (1-\sigma)\left(p \cdot a \cdot \left(2x_0^4\right)^{\frac{1}{4}} + (1-p) \cdot a \cdot (x_0^4 + x_1^4)^{\frac{1}{4}}\right) - x_0^2\right) \\ &+ (1-p) \cdot \left(\sigma \cdot a \cdot \left(x_1^4 + \left(\frac{3}{2}\right)^4\right)^{\frac{1}{4}} + (1-\sigma)\left(p \cdot a \cdot (x_0^4 + x_1^4)^{\frac{1}{4}} + (1-p) \cdot a \cdot \left(2x_1^4\right)^{\frac{1}{4}}\right) - x_1^2\right) \end{aligned}$$

With $a = 2^{2-\frac{1}{\beta}}$ and $\sigma = \frac{1}{2}$, this is

$$\begin{aligned} &p \cdot \left(\frac{2^{2-\frac{1}{4}}}{2} \cdot \left(x_0^4 + \left(\frac{3}{2}\right)^4\right)^{\frac{1}{4}} + 2 \cdot p \cdot x_0 + \frac{2^{2-\frac{1}{4}}}{2} \cdot (1-p) \cdot (x_0^4 + x_1^4)^{\frac{1}{4}} - x_0^2\right) \\ &+ (1-p) \cdot \left(\frac{2^{2-\frac{1}{4}}}{2} \cdot \left(x_1^4 + \left(\frac{3}{2}\right)^4\right)^{\frac{1}{4}} + \frac{2^{2-\frac{1}{4}}}{2} \cdot p \cdot (x_0^4 + x_1^4)^{\frac{1}{4}} + 2 \cdot (1-p) \cdot x_1 - x_1^2\right) \end{aligned}$$

For $p = \frac{1}{2}$, $x_0 = 1.45$ and $x_1 = 1.55$, this is approximately 3.75124. This is larger than $\frac{15}{4}$, which is the fitness of Homo Hamiltonensis. An entrant with these preferences that plays this equilibrium strategy therefore earns a fitness that is higher than the fitness of Homo Hamitonensis in the limit of $\epsilon \downarrow 0$.

The proof of Theorem 1 in Alger and Weibull (2013) indicates why, if the fitness of the mutant is strictly *lower* than the fitness of the resident in any equilibrium, this implies that there is an $\bar{\epsilon}$ such that the same holds for all population states with a mutant share ϵ that is smaller than $\bar{\epsilon}$. These reasons – which guarantee continuity of payoffs as ϵ changes – also imply that if there is an equilibrium for which the fitness of the mutant is strictly *higher* than that of the resident, which is the case here, there will be equilibria where the same is true, for sufficiently small ϵ. Thereby there is no $\bar{\epsilon}$ such that, if $\epsilon < \bar{\epsilon}$, the fitness of the mutant is smaller than the fitness of the resident in all equilibria. In spite of what is claimed in Theorem 1, Homo Hamiltonensis is therefore not evolutionarily stable against this mutant, while the mutant is not a behavioral alike.

For $\beta < 3$ it is not possible to create such a mutant, but for $\beta > 3$ there are mutants with a range of values for x_0, x_1 and p that can invade. In Appendix A, the second order condition for a

regular altruist with altruism parameter $\alpha = \sigma$ is derived, and there we find that it is satisfied for $\beta < 3$, but not for $\beta > 3$, while we have seen that the first order condition for Homo Hamiltonensis and regular altruist with altruism parameter $\alpha = \sigma$ is the same. In Section 5 we will see that this is not a coincidence, and that, when Homo Hamiltonensis and regular altruists differ in their equilibrium behaviour in the way they do for $\beta \in \left(3, \frac{3+\sigma}{1-\sigma}\right)$, it is actually the altruists that are stable, and Homo Hamiltonensis that becomes unstable.

Finally, it is possible to repair this relatively minor failure of the theorem. The reason why the proof overlooks this case, is that it uses their Lemma 1, which puts restrictions on the shape of the mutants, by requiring concavity, without making sure that these restrictions are satisfied at the point the lemma is invoked. That means that the theorem can be repaired by restricting the set of admissible mutants so that only those that are concave at equilibria for small ϵ remain.

4 Behavioral alikes may be able to invade Homo Hamiltonensis

The main theorem in Alger and Weibull (2013) states that Homo Hamiltonensis is evolutionarily stable against all types that are not behaviorally equivalent. The idea behind the exclusion of the behavioral alikes is that at $\epsilon = 0$, these choose the same strategy as Homo Hamiltonensis, and therefore will not be selected for or against. The definition of a behavioral alike, however, only focuses on what they do at $\epsilon = 0$, and does not look at what they do at $\epsilon > 0$. With the next example, we will show that even if a mutant preference may behave the same at $\epsilon = 0$, and therefore does not have a fitness advantage at $\epsilon = 0$, there may be equilibria for $\epsilon > 0$ in which they behave differently, and at which the behavioral alike does have a higher fitness than Homo Hamiltonensis, however small we choose ϵ.

4.1 Example 2

Consider the game

$$\pi(x, y) = 4 \cdot \min\{x, y\} - x^2$$

We assume that individuals are matched with a random draw from the population with probability $1 - \sigma$, and with a copy of themselves for sure with probability σ. This makes σ not only the assortment parameter in the limit of $\epsilon \downarrow 0$, but for all ϵ. If θ refers to the resident and τ to the mutant, then $Pr[\theta|\theta, \epsilon] = (1-\sigma)(1-\epsilon) + \sigma = 1 - (1-\sigma)\epsilon$, $Pr[\tau|\theta, \epsilon] = (1-\sigma)\epsilon$, $Pr[\theta|\tau, \epsilon] = (1-\sigma)(1-\epsilon)$, and $Pr[\tau|\tau, \epsilon] = (1-\sigma)\epsilon + \sigma$.

For a resident Homo Hamiltonensis, every strategy in $[2\sigma, 2]$ is an equilibrium strategy at

$\epsilon = 0$ (the derivation is in Appendix B). Now we can choose $\sigma = \frac{1}{2}$ and consider a mutant that has the following utility function: $u_\tau(y,x) = -(y-x)^2$ if $x \geq 1$ and $y \leq \frac{3x-1}{2}$, or if $x < 1$ and $y \geq \frac{3x-1}{2}$, and $u_\tau(y,x) = -(y-2x+1)^2$ if $x \geq 1$ and $y > \frac{3x-1}{2}$, or if $x < 1$ and $y < \frac{3x-1}{2}$. This mutant is a behavioral alike; at $\epsilon = 0$, if Homo Hamiltonensis plays $x^* = 1$, then the mutant's best response to that would be to also play $y^* = 1$, and hence $(1,1) \in B^{NE}(\sigma,\tau,0)$. Yet, if we consider equilibria at $\epsilon > 0$, then these include $(x^*_\epsilon, y^*_\epsilon) = \left(1+\frac{\epsilon}{2}, 1+\epsilon\right)$ (this derivation is also in Appendix B). For such equilibria, we will compute the material payoffs of the resident and of the mutant.

If the resident plays x^*_ϵ, and a mutant plays $y^*_\epsilon > x^*_\epsilon$, then the material payoff of the mutant is

$$\Pi_\tau(x,y,\epsilon) = (1-\sigma)(1-\epsilon)\cdot 4x^*_\epsilon + ((1-\sigma)\epsilon+\sigma)\cdot 4y^*_\epsilon - (y^*_\epsilon)^2$$

while the material payoff of the resident is

$$\Pi_\theta(x,y,\epsilon) = (1-(1-\sigma)\epsilon)\cdot 4x^*_\epsilon + (1-\sigma)\epsilon\cdot 4x^*_\epsilon - (x^*_\epsilon)^2$$

The difference between them is

$$\Pi_\tau(x,y,\epsilon) - \Pi_\theta(x,y,\epsilon) = ((1-\sigma)\epsilon+\sigma)\cdot 4\,(y^*_\epsilon - x^*_\epsilon) + (x^*_\epsilon)^2 - (y^*_\epsilon)^2$$

For $x^*_\epsilon = 1+\frac{\epsilon}{2}$, and $y^*_\epsilon = 1+\epsilon$, this is

$$\begin{aligned}\Pi_\tau(x,y,\epsilon) - \Pi_\theta(x,y,\epsilon) &= \left(\frac{\epsilon}{2}+\frac{1}{2}\right)\cdot 2\cdot\epsilon + (1+\frac{\epsilon}{2})^2 - (1+\epsilon)^2 \\ &= \epsilon^2 + \epsilon + \epsilon + \frac{\epsilon^2}{4} - 2\epsilon - \epsilon^2 = \frac{\epsilon^2}{4} > 0\end{aligned}$$

This implies that at such an equilibrium, which exists however small ϵ is, the mutant type has a fitness advantage over the resident. One can actually make intervals of equilibria for any given ϵ, all of which come with a fitness advantage of the mutant over the resident.

The reason why with strategy evolution, the existence of even a single mutant with a fitness advantage disqualifies a strategy from evolutionary stability, is that we think of evolution as a patient process. Even if there are many other mutants with a disadvantage, what matters is whether or not there (also) exists a mutant with a selective advantage. If it exists, evolution will find it. Also preference evolution we will think of as a patient process. There will be multiple mutants, and per mutant there may be multiple equilibria. For some mutant preferences, none of the equilibria will give it a fitness advantage. For other mutant preferences, some equilibria will also not give it a fitness advantage. What matters, however, is whether or not there exists

a combination of a preference and an equilibrium which gives the mutant a material payoff that is higher than the material payoff of the resident. This example shows that such a combination may exist for a mutant that is not considered, because it is classified as a behavioral alike due to its equilibrium behavior at $\epsilon = 0$.

There are three more remarks to be made regarding stability.

Behavioral alikes only need to behave alike at one equilibrium

For a strategy to fit the definition of a behavioral alike of Homo Hamiltonensis, it is enough if there is one equilibrium for which it behaves the same as Homo Hamiltonensis at $\epsilon = 0$. The definition does not require that it chooses the same strategy at every equilibrium at $\epsilon = 0$. If we do indeed consider Homo Hamiltonensis as a resident, this is not a problem. In general, however, if we consider other residents as possible candidates for stability, this can also disregard mutant preferences that on the one hand have one or more equilibria with the same material payoff at $\epsilon = 0$, and on the other equilibria that give the mutant preference a strict fitness advantage, even at $\epsilon = 0$. For obvious reasons, that would disregard mutants that could invade.

Behavioral alikes may be stepping stones for indirect invasions

In strategy evolution, even if strategies are proper neutral mutants (which means that they always earn the same payoff as the resident in any mix of the two), they may not be harmless. Neutral mutants may open doors for successive mutants, that did not have a selective advantage without the neutral mutant (see van Veelen (2012)). The same can apply to the behavioral alikes in this context, even if none of the other complications apply; they may open doors for other mutants that would not open without them.

With infinitely many strategies, there may be sequences of invasion barriers that tend to 0

With finitely many pure strategies, being evolutionarily stable against all possible mutants implies that there is also a uniform invasion barrier, or, in other words, an $\bar{\epsilon}$ that does not depend on the particular mutant, and which guarantees that every mutant will be at a disadvantage as long as its share remains below $\bar{\epsilon}$ (see Weibull (1997)). With infinitely many pure strategies, this implication does not hold, and it might be possible that there is a sequence of strategies with invasion barriers that tend to 0 (see van Veelen and Spreij (2009)). In the setting of Alger and Weibull (2013), if there are infinitely many types, the same applies. Since their setup allows for continuous strategy sets, type sets that they define as "rich" (which means that for every

strategy, there is a type for which this strategy is dominant) automatically contain infinitely many types.

5 Homo Hamiltonensis may be unstable if mixed equilibria are allowed for

In this section we want to allow for the possibility that individuals play mixed strategies, also in equilibrium. We therefore assume that individuals can randomize to break ties in their best responses. Instead of restricting them to playing pure strategies x and y, we will allow them to choose probability measures μ and ν. The definition of a (possibly mixed) Bayesian Nash equilibrium then becomes

Definition 3. *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^**

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

In the example below, we will see that if we lift the restriction on mixed equilibria, then the central result in Alger and Weibull (2013) no longer holds. In this example there are no symmetric pure equilibria that a resident Homo Hamiltonensis can play at $\epsilon = 0$, but there is a symmetric mixed equilibrium.

5.1 Example 3

Consider the game

$$\pi(x,y) = 4 \cdot \max\{x,y\} - x^2$$

and assume that the assortment parameter is $\sigma = \frac{1}{2}$.

A resident Homo Hamiltonensis can be invaded

Assume that the resident in the population is a Homo Moralis with $\kappa = \frac{1}{2}$. In Appendix we show that a uniform distribution on $[1,2]$ is a mixed BNE at $\epsilon = 0$, and that its material payoff is $\frac{13}{3}$. We also show there that if we consider a mutant that plays $z \in [0,1]$ with probability $\frac{1}{2}$, and $3-z$, also with probability $\frac{1}{2}$, then this mutant earns a material payoff of $\frac{9}{2} + z - z^2$. This makes

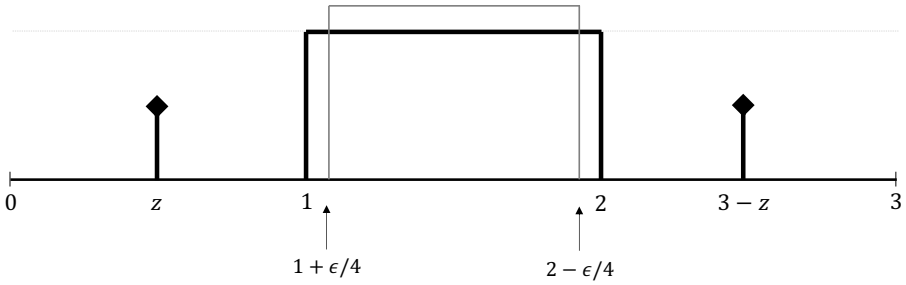

Figure 1: The equilibrium strategy for a Homo Hamiltonensis for Example 3, with $\sigma = \frac{1}{2}$, is a uniform distribution on $\left[1+\frac{\epsilon}{4}, 2-\frac{\epsilon}{4}\right]$. At $\epsilon = 0$, this gives a uniform distribution on $[0,1]$. It can be invaded by a mixed strategy that puts some mass to the left and some mass to the right of this interval.

the invasion fitness of the mutant larger than the fitness of the resident Homo Hamiltonensis for all $0 \leq z \leq 1$. In Appendix C.2 we moreover show that for $\epsilon > 0$, a uniform distribution on $\left[1+\frac{\epsilon}{4}, 2-\frac{\epsilon}{4}\right]$ is a mixed BNE, and that the material payoffs of resident and mutant change continuously as a function of ϵ. This implies that the mutant does better than the resident, at least for low mutant shares, and therefore that it can invade.

This example shows that the first part of Theorem 1 in Alger and Weibull (2013) would cease to hold if we were to allow for mixed equilibria. The second part of their theorem states that types that play strategies that would *not* maximize the utility of a Homo Hamiltonensis can be invaded, provided that the strategy space is "rich". Being rich means that for every strategy, there is a type for which this strategy is dominant. In the proof, they then show that if such a "committed" type always plays what Homo Hamiltonensis would play at $\epsilon = 0$, it can invade. Below, we will consider a resident regular altruist in order to show that this approach would also stop working, if we allow for mixed strategies.

A resident regular altruists cannot be invaded by a type that is committed to the strategy that a mutant Homo Hamiltonensis would play at $\epsilon = 0$

Assume that the resident in the population is a regular altruist with $\alpha = \frac{1}{2}$. In Appendix C.3 we show that a uniform distribution on $[0,3]$ is a mixed BNE at $\epsilon = 0$, and that its material payoff is 5. We also show there that if we consider a mutant that plays a pure strategy $y \in [0,3]$, then this mutant earns a material payoff of $3+2y-\frac{2}{3}y^2$. This makes the invasion fitness of the mutant smaller than the fitness of the resident regular altruist for all $0 \leq y \leq 3$. At $\epsilon = 0$, a mutant Homo Hamiltonensis would play $y = \frac{9}{5}$, which is between 0 and 3, and a strategy that is committed to

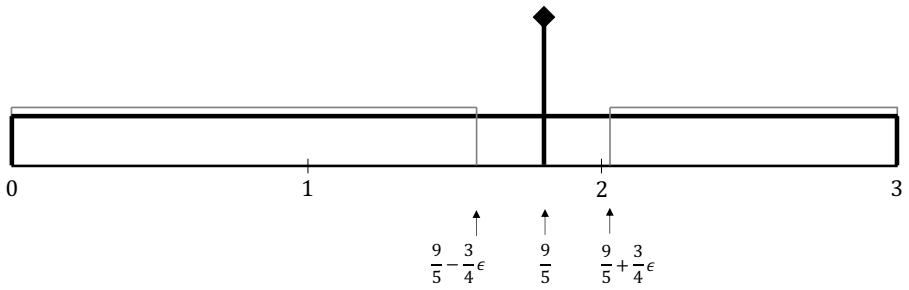

Figure 2: The equilibrium strategy for a resident regular altruist for Example 3, with $\sigma = \frac{1}{2}$, is a uniform distribution on $\left[0, z - \frac{3}{4}\epsilon\right] \cup \left[z + \frac{3}{4}\epsilon, 3\right]$, if the mutant plays z. At $\epsilon = 0$, this gives a uniform distribution on $[0, 3]$. It cannot be invaded by a strategy that is committed to playing what a mutant Homo Hamiltonesis would play at $\epsilon = 0$, which is $\frac{9}{5}$.

playing $\frac{9}{5}$ would therefore not be able to invade. In Appendix C.4 we moreover show that for $\epsilon > 0$, a uniform distribution on $\left[0, z - \frac{3}{4}\epsilon\right] \cup \left[z + \frac{3}{4}\epsilon, 3\right]$ is a mixed BNE when the mutant plays z, and that the material payoffs of resident and mutant change continuously as a function of ϵ. This implies that the resident does better than the mutant, at least for low enough mutant shares ϵ, and therefore it cannot invade.

5.2 An intuition for why Homo Moralis can sometimes be invaded when mixed equilibria are allowed for

In order to get an intuition for why the theorem ceases to hold when mixed equilibria are allowed for, we can go back to the intuition for why it did hold (under conditions) for pure equilibria. That intuition was that maximizing the utility function of Homo Hamiltonensis is the same as maximizing invasion fitness, which in equilibrium preempts possible invasions. For the more general setting, we would now also consider the possibility that playing a *mixed* equilibrium strategy is needed to preempt invasions. Facing a resident strategy ν, maximizing invasion fitness would be to choose a distribution μ that maximizes

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

If ν is a pure strategy, with a distribution that puts all mass on y, then this would be equal to choosing a distribution μ that maximizes

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

This is not the same as what Homo Hamiltonensis does, because given a match that plays y, Homo Hamiltonensis would choose an x that maximizes

$$(1-\sigma)\cdot\pi(x,y)+\sigma\cdot\pi(x,x).$$

Facing a distribution ν, Homo Hamiltonensis would choose x so as to maximize

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

If she has multiple best responses, she 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

$$(1-\sigma)\int\int\pi(x,y)d\mu(x)d\nu(y)+\sigma\int\pi(x,x)d\mu(x) \tag{2}$$

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

The difference between equations (1) and (2) is in the second term. If we think of a mutant type playing a mixed strategy, then, when two mutants meet each other, they both randomize independently, and therefore they do not necessarily play the same strategy. The second term in (1) therefore weighs the different payoffs with the probabilities that different strategy profiles occur, given two randomizing mutants, which is what it should do if the aim is to maximize the invasion fitness of the mutant. The second term in (2) on the other hand only considers symmetric strategy profiles, which is what it should do if we are looking for a (mixed) strategy that maximizes the utility of the normal Homo Hamiltonensis. It does however imply that if the second terms in (1) and (2) differ, then invasion fitness is not maximized at the equilibrium with a Homo Hamiltonensis resident, and Homo Hamiltonensis can be invaded. When strategies are pure, this difference of course disappears, and we are back in the situation where the original intuition applies. With this in mind, one could think of a generalized version of Homo Moralis, whose preferences are not functions of pure strategy profiles, but of mixed strategy profiles. While the second term for the normal Homo Moralis reflects the hypothetical payoff one would get if both would choose the same strategy, the second term for the generalized version of Homo Moralis would reflect the hypothetical expected payoff that one would get if both were to choose the same distribution. If all the weight would be on the second term – in which case Alger and Weibull (2013) refer to Homo Moralis as Homo Kantiensis – the normal Homo Moralis would choose the *strategy* that one would want to be chosen by everyone, and the generalized version would choose the *distribution* that one would want to be chosen by everyone.

To summarize and emphasize this observation, we repeat it slightly more formally.

Observation 1. *If distribution μ^* maximizes*

$$(1-\kappa)\int\int \pi(x,y)d\mu(x)d\mu^*(y) + \kappa \int\int \pi(x,y)d\mu(x)d\mu(y) \tag{3}$$

for $\kappa = \sigma$, 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. □

We will not formulate a generalized counterpart of the central result from Alger and Weibull (2013), which would involve showing under which additional conditions this would imply that for any mutant type τ that is not behaviourally equivalent to the generalized Homo Hamiltonensis, and for all equilibria between the generalized Homo Hamiltonensis and the mutant, there is an $\bar{\epsilon}$ such that the resident generalized Homo Hamiltonensis has a fitness advantage over the mutant type for all $\epsilon \in (0, \bar{\epsilon})$. The reason is that, while the problem pointed out in Section 3 disappears when looking at this more general Homo Hamiltonensis, the other complications, including the one from Section 4, remain. The problem there was that behavioral alikes, while having the same fitness at $\epsilon = 0$, may have a fitness that is higher than the fitness of the resident for ϵ close to 0. Therefore, even if for every type that is not a behavioral alike, and for every equilibrium between them, there is an interval of mutant shares $(0, \bar{\epsilon})$ such that the generalized Homo Hamiltonensis does strictly better for mutant shares ϵ in this interval, it would not necessarily be true that there will be an interval of mutant shares $(0, \bar{\epsilon})$ such that the resident never does worse than the mutant for all mutant shares ϵ within this interval. A generalization of the central result in Alger and Weibull (2013) therefore would not be dynamically meaningful. Also the problem that we will point to in Section 6, which is that Homo Hamiltonensis can be invaded if we allow for coordination on asymmetric equilibria, will remain present. It should be noted though that if we do not allow for coordination on asymmetric equilibria, but do allow for mixed equilibria, it is fair to say that this observation implies that the generalized Homo Hamiltonensis is a better candidate for stability than preferences that play equilibria in which (3) is not maximized.

5.3 Any equilibrium between generalized Homo Hamiltonensisses is also an equilibrium between regular altruists

An equally important observation is that any equilibrium μ^* for a resident generalized Homo Hamiltonensis 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 the generalized version of Homo Moralis, 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

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

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, its marginal contribution to the utility 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 $\mu^\star$ is an equilibrium for a resident generalized Homo Hamiltonensis, 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 $\alpha = \sigma$. □

Note that this is a stronger statement than can be made for the normal Homo Moralis. With pure strategies, a point that satisfies the first order condition can be an equilibrium for a resident Homo Hamiltonensis and not for regular altruists, or vice versa, depending on the difference in the second order conditions.

5.4 How these observations help

These observations allow for different types of discrepancies. It can be that an equilibrium for a resident normal Homo Hamiltonensis is not an equilibrium for a resident generalized Homo Hamiltonensis. In that case a resident Homo Hamiltonensis is unstable. This is the case for $\beta \in \left(3, \frac{3+\sigma}{1-\sigma}\right)$ in Example 1, and in Example 3. It is also no coincidence that in these examples, Homo Hamiltonensis starts disagreeing with regular altruists at the same point at which the normal Homo Hamiltonensis starts disagreeing with the generalized Homo Hamiltonensis.

The second type of discrepancy these observations allow for is that it is possible that μ^* is an equilibrium for a resident regular altruist, but not for a resident generalized Homo Hamiltonensis. At such an equilibrium, regular altruists would be vulnerable to invasions of mutants that play an action that does maximize invasion fitness. If a generalized Homo Hamiltonensis would establish herself as a resident, however, then at any equilibrium, a regular altruist playing the same equilibrium would be a proper neutral mutant, as this equilibrium would also be an equilibrium in any mix of the two.

This discrepancy between regular altruists and generalized Homo Hamitonenis can also happen if Homo Hamiltonensis and the generalized Homo Hamiltonensis do agree with each other.

Considering the game from Example 2, one can show that for a resident Homo Hamiltonensis, every $x^* \in [2\sigma, 2]$ is an equilibrium, while for a resident regular altruist with $\alpha = \sigma$, every $x^* \in [0, 2 + 2\sigma]$ is an equilibrium. The latter interval includes the former interval. For equilibria in the intersection, they are proper neutral mutants of each other. Equilibria for regular altruists that are not equilibria for Homo Hamiltonensis, however, can be invaded.

Finally, there are two more things that are worth mentioning. The first is that also if we allow for mixed strategies, as we did in Definition 3, BNE's may not exist. That should not come as a surprise, as a generalized Homo Hamiltonensis picks the (possibly 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 a generalized Homo Hamiltonensis will always be able to find a mutant strategy that can invade (see Example 3 in Bomze, Schachinger, and Weibull (2020)).

The second thing to notice is that in Section 4 of Alger and Weibull (2013), they consider finite games, in which strategies x and y are interpreted as mixed strategies over a finite set of pure strategies. That is only possible for a limited set of payoff functions $\pi(x, y)$ – they have to be bilinear – and it is different from the general setup of the rest of the paper, and from what we do here, where x and y are pure strategies, over which one could also mix.

6 Homo Hamiltonensis may be unstable when coordination on asymmetric equilibria is allowed for

The theory in Alger and Weibull (2013) does allow for a moderate degree of coordination. When there are multiple symmetric pure BNE's, they assume that everyone in the population assumes, and rightly so, that everyone else in the population plays one and the same BNE. What the current theory does not allow for is coordination on asymmetric equilibria, and it also does not allow different pairs within the population to coordinate on different equilibria, symmetric or asymmetric.

We would like to allow for coordination on asymmetric equilibria too. This implies that within every match, a randomization device is needed in order to determine who gets to play which role in the BNE – which is not needed for the coordination that was already happening in Alger and Weibull (2013). Half of the players of any type will then end up in one role, half in the other.

6.1 Example 4

We return to the game

$$\pi(x, y) = 4 \cdot \max\{x, y\} - x^2$$

and this time we will consider the asymmetric BNE's.

Equilibrium behaviour resident Homo Hamiltonensis

Assume that the resident in the population is Homo Hamiltonensis. Facing strategy y, she derives utility $(1-\sigma)\cdot 4y + \sigma \cdot 4x - x^2$ from playing $x \leq y$, and utility $4x - x^2$ from playing $x > y$. The function $(1-\sigma)\cdot 4y + \sigma \cdot 4x - x^2$ is maximized at $x = 2\sigma$, while $4x - x^2$ is maximized at $x = 2$. Which of those two is the best response depends on whether they are admissible, given the value of y, and, if both are, which results in higher utility. The best response therefore would be to choose $x = 2$ if $y \leq 1 + \sigma$, and $x = 2\sigma$ if $y > 1 + \sigma$. Between two Homo Hamiltonensisses, the asymmetric equilibria therefore are $(2\sigma, 2)$ and $(2, 2\sigma)$. The randomization device then always picks one of the players to play the high value (2) and the other to play the low value (2σ).

Equilibrium behaviour mutant regular altruist

Facing a strategy x, the utility that a regular altruist with altruism parameter σ derives from playing strategy y is $(1+\sigma)\cdot 4x - y^2$ for $y \leq x$, and $(1+\sigma)\cdot 4y - y^2$ for $y > x$. The function $(1+\sigma)\,4x - y^2$ is maximized at $y = 0$, while $(1+\sigma)\,4y - y^2$ is maximized at $y = 2\,(1+\sigma)$. Which of those two is the best response depends on whether they are admissible, given the value of x, and, if both are, which results in higher utility. The best response therefore would be to choose $y = 2(1+\sigma)$ if $x \leq 1 + \sigma$, and $y = 0$ if $x > 1 + \sigma$.

The mutant regular altruists can use the same coordination device as Homo Hamiltonensis does when they meet each other, but play different values when they are picked to play the high or the low value, respectively. This way, mutant altruists end up playing $(0, 2)$ and $(2 + 2\sigma, 2\sigma)$ when matched with Homo Hamiltonensis, and $(0, 2 + 2\sigma)$ and $(2 + 2\sigma, 0)$ with themselves. What makes this example easy, is that, if the other player plays the high value (2 when the other is a Homo Hamiltonensis, and $2 + 2\sigma$ when the other is a regular altruist), in either case the best response of a regular altruist is to play 0. Similarly, if the other player plays the low value, in either case it is a best response for a regular altruist to play $2 + 2\sigma$.

Material payoffs

Players will be equally likely to be picked to play the high or the low value. The material payoff of a resident Homo Hamiltonensis is

$$\frac{1}{2}\left(8-(2\sigma)^2\right)+\frac{1}{2}\left(8-2^2\right)=6-2\sigma^2$$

The invasion fitness of a mutant altruist, with altruism parameter $\alpha=\sigma$, is

$$\begin{aligned}&(1-\sigma)\left(\frac{1}{2}\left(4\cdot 2-0^2\right)+\frac{1}{2}\left(4\cdot(2+2\sigma)-(2+2\sigma)^2\right)\right)\\&\quad+\sigma\left(\frac{1}{2}\left(4\cdot(2+2\sigma)-0^2\right)+\frac{1}{2}\left(4\cdot(2+2\sigma)-(2+2\sigma)^2\right)\right)\\&=8+8\sigma-(1-\sigma)4\sigma-\frac{1}{2}(2+2\sigma)^2=8+4\sigma+4\sigma^2-2-4\sigma-2\sigma^2=6+2\sigma^2.\end{aligned}$$

A mutant regular altruist with altruism parameter $\alpha=\sigma$ therefore can invade Homo Hamiltonensis.

Equilibrium behaviour resident regular altruists

Assume that the resident in the population is an altruist with altruism parameter σ. Their best response, as calculated above, would be to choose $x=0$ if $y\geq 1+\sigma$, and $x=2(1+\sigma)$ if $y<1+\sigma$, leading to equilibria $(0,2+2\sigma)$ and $(2+2\sigma,0)$.

Equilibrium behaviour mutant Homo Hamiltonensis

Assume that the resident in the population is a regular altruist with altruism parameter σ. The best response of Homo Hamiltonensis, as computed above, is $x=2$ if $y\leq 1+\sigma$, and $x=2\sigma$ if $y>1+\sigma$. Given that these result in the same best responses against both types, when the other is picked to play high or low, respectively, they are also the same against any probability distribution over those types, and we will see mutant Homo Hamiltonensis play $(2\sigma,2+2\sigma)$ and $(2,0)$ when matched with resident regular altruists, and $(2\sigma,2)$ and $(2,2\sigma)$ when matched with copies of themselves.

Material payoffs

Players will be equally likely to be picked to play the high or the low value. The material payoff of a resident regular altruist therefore becomes

$$\begin{aligned}&\frac{1}{2}\left(8+8\sigma-0^2\right)+\frac{1}{2}\left(8+8\sigma-(2+2\sigma)^2\right)\\&=8+8\sigma-2-4\sigma-2\sigma^2=6+4\sigma-2\sigma^2\end{aligned}$$

The invasion fitness of a mutant Homo Hamiltonensis is

$$(1-\sigma)\left(\frac{1}{2}\left(4\cdot(2+2\sigma)-(2\sigma)^2\right)+\frac{1}{2}\left(4\cdot 2-2^2\right)\right)+\sigma\left(\frac{1}{2}\left(4\cdot 2-(2\sigma)^2\right)+\frac{1}{2}\left(4\cdot 2-2^2\right)\right)$$
$$=8+(1-\sigma)4\sigma-\frac{1}{2}(2\sigma)^2-\frac{1}{2}2^2=8+4\sigma-4\sigma^2-2\sigma^2-2=6+4\sigma-6\sigma^2.$$

A mutant Homo Hamiltonensis, or a type that is committed to what a mutant Homo Hamiltonensis would do at $\epsilon = 0$, therefore cannot invade a regular altruist with altruism parameter $\alpha = \sigma$.

Reducing waste

It may be worth observing that both allowing for mixed strategies, as we did in Section 5, and allowing for coordination on asymmetric equilibria, as we did in this section, have the effect that they reduce, or stop, material payoff from being "wasted" on imaginary payoffs. It would not be strange to expect that evolution could be creative in finding ways not to have to sacrifice material payoffs for hypothetical payoffs, when these hypothetical payoffs never materialize. In these cases, it is also not always clear that maximizing the utility function of a Homo Kantiensis would be the morally right thing to do. That would be particularly hard to defend in examples where increases in those hypothetical payoffs come at the expense, both of the material payoff of the agent herself, and of the material payoff of the individual she is matched with.

7 Cross-species empirical evidence

In this section, we would like to look at the broader, cross-species empirical evidence. The core idea of the central result in Alger and Weibull (2013) is that human morality has evolved as a consequence of a combination of incomplete information about the preferences of the other, and assortment in the population. Morality here is defined as a preference that puts a positive weight on doing what you would wish everyone would do. We will consider both assortment and incomplete information, and ask ourselves if humans would be the first species in which we would expect morality to evolve, if the key lies in this combination of ingredients.

Assortment, the cancellation effect, and morality towards unrelated individuals

Alger and Weibull (2013) consider two mechanisms that can create assortment: (1) interactions between kin, and (2) living in a group structured population. The interactions between kin that they describe require kin recognition. In this case, behaviour towards for instance siblings would be more moral than behaviour towards random others. Humans are not unique in their ability to

recognize kin (Lieberman, Tooby, and Cosmides, 2007), or otherwise in having the opportunity to behave differently towards kin than to non-kin; see for instance Bergman, Beehner, Cheney, and Seyfarth (2003) on the ability of baboons to differentiate between within-family and between-family interactions, or Silk (2009) for a review of kin biases in non-human primates. Compared to for instance chimpanzees, humans do however have a larger repertoire of kin they differentiate their behaviour towards. Langergraber, Mitani, and Vigilant (2007) show that male chimpanzees affiliate and cooperate more with maternal brothers, but not with paternal brothers, than they do with non-kin. The results for paternal brothers are likely due to unreliable recognition of paternal sibship in chimpanzees (see also Parr and de Waal (1999) on the limited ability of chimpanzees in visual kin recognition).

Living in a group structured population can imply that individuals within the same group end up positively assorted. With genetic transmission, that makes them kin, but cultural transmission can have the same effect. Alger and Weibull (2013) also consider the possibility that a group structure induces group members to have the same preferences because of a shared shock within the group. There is no reason to believe that the group structure in ancestral human populations was special with respect to genetic relatedness. Studies that compare humans with non-human primates for instance find comparable levels of genetic differentiation (Fischer, Pollack, Thalmann, Nickel, and Pääbo (2006); Langergraber, Schubert, Rowney, Wrangham, Zommers, and Vigilant (2011); Scally, Yngvadottir, Xue, Ayub, Durbin, and Tyler-Smith (2013)). Moreover, modern conditions make humans interact with relatively many unrelated individuals, which suggests low levels of genetic assortment in most of our daily interactions, while human morality also pertains to behaviour towards non-related others. Bell, Richerson, and McElreath (2009) do however find that the cultural differentiation between human groups is larger than the genetic differentiation. This is hard to compare with other species, but it could be a way in which humans are special.

There is one more thing that is worth mentioning here. The interactions in which the games in Alger and Weibull (2013) are played, happen in an assorted way. For this to imply that those that get a higher payoff will always outcompete those with a lower payoff, it would have to be assumed that competition happens in a non-assorted, or less assorted way. If competition also takes place in an assorted way, then the *cancellation effect* occurs (Wilson, Pollock, and Dugatkin (1992), Taylor (1992b,a), see also van Veelen, Allen, Hoffman, Simon, and Veller (2017) for how this relates to Hamilton's rule). Many population structures induce assortment, both with respect to the games that offer possibilities for cooperation, and with respect to competition. If these games happen as assortedly as competition does, then having a high payoff yourself will positively correlate with competing against individuals that also have high payoffs, and no deviations from

selfishness will evolve. The key to the evolution of pro-social behaviour by assortment therefore is not that interactions happen in an assorted way, but that there is a discrepancy between, on the one hand, how assorted opportunities for cooperation are, and, on the other, how assorted competition is. This is also the reason why kin recognition works, because that is a mechanism that can create such a discrepancy. The cancellation effect at the individual level would apply to modern conditions, while for ancestral conditions the cancellation effect at the group level, which makes the evolution of pro-social behaviour harder, but not impossible, might be more relevant (Akdeniz and van Veelen (2020)).

Complete vs. incomplete information

One may want to contrast a setting with assortment and incomplete information to a setting with assortment and complete information. For the latter, Alger and Weibull (2012) show that altruism can evolve. The degree of altruism is determined, not only by the level of assortment in the population, but also by the degree to which the game has strategic complements or substitutes. Alger and Weibull (2012) show that the presence of strategic complements increases the equilibrium level of altruism, because being altruistic allows individuals to commit to playing nice, and strategic substitutes decrease the equilibrium level of altruism, because being less altruistic, or even spiteful, allows individuals to commit to playing nasty.[2] There is however one reason why these papers are not perfectly comparable, and one reason why they might not be comparable in a cross-species context. The first is that Alger and Weibull (2012) restrict attention to preferences that are altruistic to different degrees, including selfishness and spite, and that Homo Moralis therefore is not considered. This implies that they also do not compare the stability of regular altruists and Homo Moralis. The complete versus incomplete information part might also not lend itself to a cross-species comparison, because one might argue that in both cases the equilibrium play requires cognitive abilities that restricts both of those models to humans anyway. If one would nonetheless want to distill the central messages from Alger and Weibull (2012) and Alger and Weibull (2013) and put them side to side, then that would be that, when combined with population structure, *complete* information about the preferences of the other leads to altruism, and *incomplete* information to morality. Alternatively, one can not aim at making a comparison, and just focus on the incomplete information from Alger and Weibull (2013), and ask to what extent this matches humans.

If we compare Homo Sapiens with other species, then humans are special in the degree to

[2] Another setting where altruistic preferences can evolve, is one where interactions are not really games, but simply encounters in which one player can choose to confer, or not to confer, a fitness benefit to another player, at a fitness cost to him or herself. Also in such a setting, with assortment, altruism evolves, and not morality, and here, in the absence of strategic substitutes or complements, it does to the same degree as the assortment in the population; see van Veelen (2006).

which they are informed about, and understand, the intentions and motives of their conspecifics (or, in terms of this result: their preferences). Although there are species that have *Theory of Mind* to some degree, there is no doubt that humans are extremely good at understanding what others want (see Call and Tomasello (2008) for a review of *Theory of Mind* in chimpanzees, Krupenye and Call (2019) for a more recent review in a wider range of species, and Penn, Holyoak, and Povinelli (2008) on the discontinuity between human and nonhuman minds). It is therefore not unreasonable to argue that for many of the behaviours in which morality plays a role, our species is closer to complete than it is to incomplete information about each other's preferences. Especially with respect to our relatives and others we regularly interact with, we tend to recognize individual differences in preferences, rather than behaving the same towards everyone within a category (such as siblings or cousins) and form a belief about what their preferences could be, using Bayes rule, what our own preferences are, and what the average or predominant preferences in the population are.

If morality evolves as a result of a combination of assortment in the population and incomplete information about what others want, humans therefore may not be the first species where one would expect to find it. Our morality extends to unrelated others, with whom assortment is zero, and our species is closer to complete, rather than incomplete information about each other's preferences, in particular those of our relatives and others we regularly interact with.

We would also like to stress that we do not argue that Homo Moralis does not exist (Miettinen, Kosfeld, Fehr, and Weibull (2020)). We only suggest that the model introduced in Alger and Weibull (2013) does not offer an evolutionary explanation that matches the cross-species comparative statics particularly well, while there are also other explanations (Frank (1987, 1988)).

A Example 1

Consider the game

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

with $a = 2^{2-\frac{1}{\beta}}$. The first order condition in a symmetric pure equilibrium, both for Homo Moralis and for regular altruists (see Section 2.2), is:

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

where π_1 and π_2 are the derivatives to the first and second argument of the payoff function. For this game, the first order condition therefore is that

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

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

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

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

For the second order conditions of this game, we first compute the following second derivatives.

$$\begin{aligned}
\pi_{1,1}(x,x) &= \left[a \cdot (\beta-1)x^{\beta-2}(x^\beta + y^\beta)^{\frac{1}{\beta}-1} + a \cdot x^{\beta-1}\beta x^{\beta-1}(\frac{1}{\beta}-1)(x^\beta + y^\beta)^{\frac{1}{\beta}-2} - 2\right]_{y=x} \\
&= a \cdot (\beta-1)x^{\beta-2}(2x^\beta)^{\frac{1}{\beta}-1} + a \cdot x^{2\beta-2}(1-\beta)(2x^\beta)^{\frac{1}{\beta}-2} - 2 \\
&= a \cdot (\beta-1)x^{-1} \cdot 2^{\frac{1}{\beta}-1} + a \cdot x^{-1}(1-\beta) \cdot 2^{\frac{1}{\beta}-2} - 2 \\
&= 2 \cdot (\beta-1)x^{-1} + x^{-1}(1-\beta) - 2 = (\beta-1)x^{-1} - 2
\end{aligned}$$

At the one before last step, we use $a = 2^{2-\frac{1}{\beta}}$.

$$\begin{aligned}
\pi_{1,2}(x,x) &= \left[a \cdot x^{\beta-1}\beta y^{\beta-1}(\frac{1}{\beta}-1)(x^\beta + y^\beta)^{\frac{1}{\beta}-2}\right]_{y=x} \\
&= a \cdot x^{2\beta-2}(1-\beta)(2x^\beta)^{\frac{1}{\beta}-2} \\
&= a \cdot (1-\beta) \cdot x^{-1} \cdot 2^{\frac{1}{\beta}-2} = (1-\beta)x^{-1}
\end{aligned}$$

Due to the symmetry of the first term in the payoff function, we can immediately see that $\pi_{1,1} = \pi_{2,2} - 2$, and hence

$$\pi_{2,2}(x,x) = (\beta-1)x^{-1}$$

These second order derivatives make the second order condition for the Homo Moralis (also from Section 2.2) for this game

$$\begin{aligned}
\pi_{1,1}(x,x) + \kappa \cdot (2\pi_{1,2}(x,x) + \pi_{2,2}(x,x)) &< 0 \\
(\beta-1)x^{-1} - 2 + \kappa\left(2(1-\beta)x^{-1} + (\beta-1)x^{-1}\right) &< 0 \\
(1-\kappa)(\beta-1)x^{-1} &< 2
\end{aligned}$$

If this is to hold at $x = x^* = 1 + \kappa$, then

$$(1-\kappa)(\beta-1) < 2(1+\kappa) \Longleftrightarrow \beta < \frac{3+\kappa}{1-\kappa}$$

The second order condition for regular altruists with $\alpha = \kappa$, on the other hand, would be

$$\begin{aligned} \pi_{1,1}(x,x) + \alpha \cdot \pi_{2,2}(x,x) &< 0 \\ (\beta-1)x^{-1} - 2 + \alpha \cdot (\beta-1)x^{-1} &< 0 \\ (1+\alpha)(\beta-1)x^{-1} &< 2 \end{aligned}$$

If this is to hold at $x = x^* = 1 + \alpha$, then

$$(\beta-1) < 2 \Longleftrightarrow \beta < 3$$

B Example 2 with $\epsilon > 0$

Consider the game

$$\pi(x,y) = 4 \cdot \min\{x,y\} - x^2$$

We assume that individuals are matched with a random draw from the population with probability $1-\sigma$, and with probability σ they are matched with a copy of themselves for sure. This makes σ not only the assortment parameter in the limit of $\epsilon \downarrow 0$, but for all ϵ. If θ refers to the resident and τ to the mutant, then that makes $Pr[\theta|\theta,\epsilon] = (1-\sigma)(1-\epsilon) + \sigma = 1 - (1-\sigma)\epsilon$, $Pr[\tau|\theta,\epsilon] = (1-\sigma)\epsilon$, $Pr[\theta|\tau,\epsilon] = (1-\sigma)(1-\epsilon)$, and $Pr[\tau|\tau,\epsilon] = (1-\sigma)\epsilon + \sigma = \sigma + \epsilon - \epsilon\sigma$.

Equilibrium behaviour resident Homo Hamiltonensis at $\epsilon > 0$

Assume that the resident in the population is a Homo Hamiltonensis, and that, together with the mutant, they play a pure, symmetric BNE $(x_\epsilon^*, y_\epsilon^*)$ for $\epsilon > 0$.

Case 1: $x_\epsilon^* < y_\epsilon^*$.

If $x \leq x_\epsilon^* < y_\epsilon^*$, Homo Moralis would have utility

$$(1-\kappa) \cdot ((1-(1-\sigma)\epsilon) \cdot 4x + (1-\sigma)\epsilon \cdot 4x) + \kappa \cdot 4x - x^2$$

which amounts to $4x - x^2$. If $x^*_\epsilon < x \leq y^*_\epsilon$, Homo Moralis would have utility

$$(1-\kappa)\cdot((1-(1-\sigma)\epsilon)\cdot 4x^*_\epsilon + (1-\sigma)\epsilon\cdot 4x) + \kappa\cdot 4x - x^2$$

If $x^*_\epsilon < y^*_\epsilon \leq x$, Homo Moralis would have utility

$$(1-\kappa)\cdot((1-(1-\sigma)\epsilon)\cdot 4x^*_\epsilon + (1-\sigma)\epsilon\cdot 4y^*_\epsilon) + \kappa\cdot 4x - x^2$$

For this utility to be maximized at $x = x^*_\epsilon$, we need $x^*_\epsilon \in [2\left((1-\kappa)(1-\sigma)\epsilon + \kappa\right), 2]$.

Case 2: $x^*_\epsilon = y^*_\epsilon$

If $x \leq x^*_\epsilon = y^*_\epsilon$, Homo Moralis has utility

$$(1-\kappa)\cdot 4x + \kappa\cdot 4x - x^2$$

which amounts to $4x - (x)^2$. If $x > x^*_\epsilon = y^*_\epsilon$, Homo Moralis has utility

$$(1-\kappa)\cdot 4x^*_\epsilon + \kappa\cdot 4x - x^2$$

For this to be maximized at $x = x^*_\epsilon = y^*_\epsilon$, we need $x^*_\epsilon \in [2\kappa, 2]$.

Case 3: $x^*_\epsilon > y^*_\epsilon$

If $x \leq y^*_\epsilon < x^*_\epsilon$, Homo Moralis has utility

$$(1-\kappa)\cdot((1-(1-\sigma)\epsilon)\cdot 4x + (1-\sigma)\epsilon\cdot 4x) + \kappa\cdot 4x - x^2$$

which amounts to $4x - x^2$. If $y^*_\epsilon < x \leq x^*_\epsilon$, Homo Moralis would have utility

$$(1-\kappa)\cdot((1-(1-\sigma)\epsilon)\cdot 4x + (1-\sigma)\epsilon\cdot 4y^*_\epsilon) + \kappa\cdot 4x - x^2$$

If $y^*_\epsilon < x^*_\epsilon < x$, Homo Moralis would have utility

$$(1-\kappa)\cdot((1-(1-\sigma)\epsilon)\cdot 4x^*_\epsilon + (1-\sigma)\epsilon\cdot 4y^*_\epsilon) + \kappa\cdot 4x - x^2$$

For this to be maximized at $x = x^*_\epsilon$, we need $x^*_\epsilon \in [2\kappa, 2\left(1-(1-\kappa)(1-\sigma)\epsilon\right))]$.

With $\kappa = \sigma$, these conditions on $(x^*_\epsilon, y^*_\epsilon)$ so that x^*_ϵ is Homo Moralis' best response to y^*_ϵ amount

to

Case 1:	$x_\epsilon^* \in \left[2\left((1-\sigma)^2\epsilon+\sigma\right),2\right]$	for $y_\epsilon^* > x_\epsilon^*$
Case 2:	$x_\epsilon^* \in [2\sigma,2]$	for $y_\epsilon^* = x_\epsilon^*$
Case 3:	$x_\epsilon^* \in \left[2\sigma,2\left(1-(1-\sigma)^2\epsilon\right)\right]$	for $y_\epsilon^* < x_\epsilon^*$

At $\epsilon = 0$, these intervals all collapse to $[2\sigma, 2]$, which makes perfect sense, since at $\epsilon = 0$ the behaviour of the mutant does not matter, and the condition for Case 2, where mutant and resident play the same strategy, applies.

Equilibrium behaviour mutant at $\sigma = \frac{1}{2}$ and $\epsilon > 0$

For the example in Section 4, we will take $\sigma = \frac{1}{2}$. If we take $x_\epsilon^* = 1 + \frac{\epsilon}{2}$, then that would make x_ϵ^* equal to the left boundary of the interval for Case 1, which assumes that $y_\epsilon^* > x_\epsilon^*$. Now consider a mutant with the following utility function: $u_\tau(y,x) = -(y-x)^2$ if $x \geq 1$ and $y \leq \frac{3x-1}{2}$, or $x < 1$ and $y \geq \frac{3x-1}{2}$, and $u_\tau(y,x) = -(y-2x+1)^2$ if $x \geq 1$ and $y > \frac{3x-1}{2}$, or $x < 1$ and $y < \frac{3x-1}{2}$, and assume that the mutant plays $y_\epsilon^* = 1 + \epsilon > 1 + \frac{\epsilon}{2} = x_\epsilon^*$. This mutant would then have utility

$$\begin{aligned}
&(1-\sigma)(1-\epsilon)\cdot u_\tau(y,x_\epsilon^*) + ((1-\sigma)\epsilon+\sigma)\cdot u_\tau(y,y_\epsilon^*) \\
=&(1-\sigma)(1-\epsilon)\left[-(y-2(1+\frac{\epsilon}{2})+1)^2\right] + ((1-\sigma)\epsilon+\sigma)\left[-\left(y-(1+\epsilon)\right)^2\right] \\
=&-\left(y-1-\epsilon\right)^2
\end{aligned}$$

for choices of y close to y_ϵ^*. Setting the derivative to y equal to 0 implies that $y = y_\epsilon^* = 1 + \epsilon$ is indeed the best response.

C Example 3

Consider the game

$$\pi(x,y) = 4 \cdot \max\{x,y\} - x^2$$

C.1 Payoffs at $\epsilon = 0$ with Homo Hamiltonensis as a resident

Equilibrium behaviour resident Homo Hamiltonensis

Assume that the resident in the population is a Homo Moralis with $\kappa = \frac{1}{2}$. At $\epsilon = 0$, where the resident only meets other residents, their utility of playing $1 \leq x \leq 2$ against a uniform

distribution on $[1,2]$ is

$$\begin{aligned}
&(1-\kappa)\int_1^2 \pi(x,y)dy + \kappa\pi(x,x) \\
&= \frac{1}{2}\left[\int_1^x \left(4x-x^2\right)dy + \int_x^2 \left(4y-x^2\right)dy\right] + \frac{1}{2}(4x-x^2) \\
&= 2\left[\int_1^x xdy + \int_x^2 ydy\right] + 2x - x^2 = 2\left[x(x-1) + \frac{1}{2}(4-x^2)\right] + 2x - x^2 = 4
\end{aligned}$$

This is independent of x. For $x > 2$, their utility is $(1-\kappa)(4x-x^2)+\kappa(4x-x^2) = 4x-x^2$, which is less than 4, and for $x < 1$, their utility is $(1-\kappa)\int_1^2 \left(4y-x^2\right)dy + \kappa(4x-x^2) = \frac{1}{2}\left[2y^2\right]_1^2 + 2x - x^2 = 3 + 2x - x^2$, which is also less than 4. A Homo Hamiltonensis therefore would not deviate from this distribution of strategies.

Material payoff resident Homo Hamiltonensis

The material payoff of a Homo Hamiltonensis is

$$\begin{aligned}
&\int_1^2\left[\int_1^2 4\cdot\max(x,y)-x^2\right]dydx = 2\cdot 4\cdot\int_1^2\left[\int_1^x x\right]dydx - \int_1^2\int_1^2 x^2dydx \\
&= 8\int_1^2 x\left(x-1\right)dx - \int_1^2 x^2dx = 7\int_1^2 x^2dx - 8\int_1^2 xdx = \frac{49}{3} - 12 = \frac{13}{3}
\end{aligned}$$

Material payoff mutant

If a mutant plays $z \in [0,1]$ with probability $\frac{1}{2}$, and $3-z$, also with probability $\frac{1}{2}$, then this mutant earns a material payoff of

$$\begin{aligned}
&= \frac{1}{2}\left((1-\sigma)\int_1^2 4xdx + \sigma(\frac{1}{2}\cdot 4z + \frac{1}{2}\cdot 4(3-z)) - z^2\right) + \frac{1}{2}\left(4(3-z) - (3-z)^2\right) \\
&= \frac{1}{2}\left((1-\sigma)\left[2x^2\right]_1^2 + 6\sigma - z^2\right) + \frac{1}{2}\left(3 + 2z - z^2\right) \\
&= \frac{1}{2}\left(6(1-\sigma) + 6\sigma - z^2\right) + \frac{1}{2}\left(3 + 2z - z^2\right) = \frac{9}{2} + z - z^2
\end{aligned}$$

This payoff, or invasion fitness, is larger than the payoff of the resident Homo Moralis for all $0 \le z \le 1$.

C.2 Payoffs at $\epsilon > 0$ with Homo Hamiltonensis as a resident

In order to see that this mutant can really invade, we still have to show that the material payoff of the mutant is not only higher than that of the resident at $\epsilon = 0$, but also that there is an

interval $(0, \bar{\epsilon})$ such that the material payoff of the mutant exceeds the material payoff of the resident for all mutant shares $\epsilon \in (0, \bar{\epsilon})$.

Equilibrium behaviour resident Homo Hamiltonensis for $\epsilon > 0$

We assume that individuals are matched with a random draw from the population with probability $1-\sigma$, and with a copy of themselves for sure with probability σ. This makes σ not only the assortment parameter in the limit of $\epsilon \downarrow 0$, but for all ϵ. If θ refers to the resident and τ to the mutant, then that makes $Pr[\theta|\theta,\epsilon] = (1-\sigma)(1-\epsilon)+\sigma = 1-\epsilon+\epsilon\sigma$, $Pr[\theta|\tau,\epsilon] = (1-\sigma)(1-\epsilon)$, $Pr[\tau|\theta,\epsilon] = (1-\sigma)\epsilon$ and $Pr[\tau|\tau,\epsilon] = (1-\sigma)\epsilon+\sigma = \sigma+\epsilon-\epsilon\sigma$.

Here we assume that $\sigma = \frac{1}{2}$, and that the resident is a Homo Moralis with morality parameter $\kappa = \sigma = \frac{1}{2}$. With a mutant that plays $z < 1$ with probability $\frac{1}{2}$, and $3-z > 2$, also with probability $\frac{1}{2}$, Homo Hamiltonensis will play a uniform distribution on $\left[1+\frac{\epsilon}{4}, 2-\frac{\epsilon}{4}\right]$, denoted with F_ϵ, in the unique, mixed BNE; the utility that Homo Hamiltonensis derives from playing $x \in \left[1+\frac{\epsilon}{4}, 2-\frac{\epsilon}{4}\right]$ in this population is

$$
\begin{aligned}
&(1-\kappa)\left((1-\epsilon+\epsilon\sigma)\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} \pi(x,y)dF_\epsilon(y) + (1-\sigma)\epsilon\left(\frac{1}{2}\pi(x,z)+\frac{1}{2}\pi(x,3-z)\right)\right) \\
&\qquad\qquad + \kappa\pi(x,x) \\
&= \frac{1}{2}\left(\left(1-\frac{\epsilon}{2}\right)\left(\int_{1+\frac{\epsilon}{2}}^{x} \frac{1}{1-\epsilon/2}\cdot 4x dy + \int_{x}^{2-\frac{\epsilon}{2}} \frac{1}{1-\epsilon/2}\cdot 4y dy\right) + \frac{\epsilon}{2}\left(\frac{1}{2}(4x)+\frac{1}{2}\cdot 4(3-z)\right)\right) \\
&\qquad\qquad + \frac{1}{2}(4x) - x^2 \\
&= \int_{1+\frac{\epsilon}{4}}^{x} 2x dy + \int_{x}^{2-\frac{\epsilon}{4}} 2y dy + \frac{\epsilon}{2}\cdot x + \frac{\epsilon}{2}\cdot(3-z) + 2x - x^2 \\
&= 2x(x-1-\frac{\epsilon}{4}) + (2-\frac{\epsilon}{4})^2 - x^2 + \frac{\epsilon}{2}\cdot x + \frac{\epsilon}{2}\cdot(3-z) + 2x - x^2 \\
&= (2-\frac{\epsilon}{4})^2 + \frac{\epsilon}{2}\cdot(3-z)
\end{aligned}
$$

This is independent of x. Utilities for values of x outside this interval are lower, so a Homo Hamiltonensis would not deviate from this distribution of strategies.

Material payoff resident Homo Hamiltonensis at $\epsilon > 0$

Here we will compute the material payoff of a resident Homo Hamiltonensis θ playing this mixed BNE in a population with assortment parameter $\sigma = \frac{1}{2}$, and with mutant τ as described above. We denote the actions in this BNE played by the resident by $x^*_{\theta,\tau,\epsilon}$ and the action played by the mutant by $y^*_{\theta,\tau,\epsilon}$, respectively, and while τ is a committed type, whose action does not depend

on ϵ, the equilibrium behaviour of resident θ does change with ϵ.

$$\begin{aligned}
\Pi_\theta(x^*_{\theta,\tau,\epsilon}, y^*_{\theta,\tau,\epsilon}, \epsilon) &= (1-\epsilon+\epsilon\sigma)\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}}\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} \frac{1}{1-\epsilon/2}\cdot 4\max(x,y)dydx \\
&\quad + (1-\sigma)\epsilon\left(\frac{1}{2}\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} \frac{1}{1-\epsilon/2}\cdot 4xdx + \frac{1}{2}\cdot 4(3-z)\right) - \int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}}\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} \frac{1}{1-\epsilon/2}\cdot x^2 dydx \\
&= \int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}}\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} 4\max(x,y)dydx \\
&\quad + \frac{\epsilon}{2}\left(\frac{1}{2}\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} \frac{1}{1-\epsilon/2}\cdot 4xdx + \frac{1}{2}\cdot 4(3-z)\right) - \int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} x^2 dx \\
&= 2\cdot 4\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}}\int_{1+\frac{\epsilon}{4}}^{x} xdydx + \epsilon\left(\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} \frac{1}{1-\epsilon/2}\cdot xdx + (3-z)\right) - \int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} x^2 dx \\
&= 8\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} x\left(x-1-\frac{\epsilon}{4}\right)dx + \epsilon\left(\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} \frac{1}{1-\epsilon/2}\cdot xdx + (3-z)\right) - \int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} x^2 dx \\
&= 8\left[\frac{1}{3}x^3 - \frac{1}{2+\epsilon/2}x^2\right]_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} + \epsilon\left(\left[\frac{1}{2-\epsilon}x^2\right]_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} + (3-z)\right) - \left[\frac{1}{3}x^3\right]_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}}
\end{aligned}$$

This is continuous in ϵ, and for $\epsilon \downarrow 0$, this converges to the material payoff of the resident Homo Hamiltonensis at $\epsilon = 0$, as computed in Appendix

$$\begin{aligned}
\lim_{\epsilon\downarrow 0}\Pi_\theta(x^*_{\theta,\tau,\epsilon}, y^*_{\theta,\tau,\epsilon}, \epsilon) &= 8\left(\frac{1}{3}(8-1) - \frac{1}{2}(4-1)\right) - \left(\frac{1}{3}(8-1)\right) \\
&= \frac{49}{3} - 12 = \frac{13}{3} = \Pi_\theta
\end{aligned}$$

Material payoff mutant at $\epsilon > 0$

The mutant, playing $0 \le z \le 1$ with probability $\frac{1}{2}$, and $2 \le 3-z \le 3$, also with probability $\frac{1}{2}$, earns a material payoff of

$$\begin{aligned}
&\Pi_\tau(x^*_{\theta,\tau,\epsilon}, y^*_{\theta,\tau,\epsilon}, \epsilon) \\
&= \frac{1}{2}\left((1-\sigma)(1-\epsilon)\int_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} \frac{1}{1-\epsilon/2}\cdot 4xdx + ((1-\sigma)\epsilon+\sigma)\left(\frac{1}{2}\cdot 4z + \frac{1}{2}\cdot 4(3-z)\right) - z^2\right) \\
&\quad + \frac{1}{2}\left(4(3-z) - (3-z)^2\right) \\
&= \frac{1}{2}\left((1-\sigma)(1-\epsilon)\left[\frac{2}{1-\epsilon/2}x^2\right]_{1+\frac{\epsilon}{4}}^{2-\frac{\epsilon}{4}} + 6((1-\sigma)\epsilon+\sigma) - z^2\right) + \frac{1}{2}\left(3+2z-z^2\right)
\end{aligned}$$

This is also continuous in ϵ, and for $\epsilon \downarrow 0$, this converges to the material payoff of the mutant at $\epsilon = 0$, as computed in Appendix **??**=[Payoffs at epsilon=0$_{\epsilon\downarrow 0}\Pi_\tau(x^*_{\theta,\tau,\epsilon}, y^*_{\theta,\tau,\epsilon}, \epsilon)$ = $\frac{1}{2}\left(6(1-\sigma)+6\sigma-z^2\right)$+ $\frac{1}{2}\left(3+2z-z^2\right)$

$$= \frac{9}{2} + z - z^2 = \Pi_{\tau,\theta}$$

In Appendix we already concluded that for all $0 \leq z \leq 1$, the invasion fitness of the mutant at $\epsilon = 0$, $\Pi_{\tau,\theta}$, is larger than the fitness of the resident Homo Hamiltonensis at $\epsilon = 0$, Π_θ. With continuity of the payoff functions, we can now also conclude that the material payoff of the mutant is larger than that of the resident for ϵ sufficiently close to 0. Homo Hamiltonensis therefore is not evolutionarily stable for this example, where we allowed for mixed BNE.

C.3 Payoffs at $\epsilon = 0$ with a regular altruist as a resident

Equilibrium behaviour resident regular altruists

Assume that the resident in the population is a regular altruist with $\alpha = \frac{1}{2}$. At $\epsilon = 0$, where the resident only meets other residents, their utility of playing $0 \leq x \leq 3$ against a uniform distribution on $[0, 3]$ is

$$\begin{aligned}
&\int_0^3 \frac{1}{3}\pi(x,y)dy + \alpha \int_0^3 \frac{1}{3}\pi(y,x)dy \\
&= \int_0^x \frac{1}{3}(4x - x^2)dy + \int_x^3 \frac{1}{3}(4y - x^2)dy + \frac{1}{2}\left(\int_0^x \frac{1}{3}(4x - y^2)dy + \int_x^3 \frac{1}{3}(4y - y^2)dy\right) \\
&= \frac{3}{2}\left(\int_0^x \frac{4}{3}xdy + \int_x^3 \frac{4}{3}ydy\right) - x^2 - \frac{1}{2}\int_0^3 \frac{1}{3}y^2dy = 2\left(x^2 + \frac{1}{2}(3^2 - x^2)\right) - x^2 - \frac{1}{6}\frac{1}{3}3^2 = \frac{17}{2}
\end{aligned}$$

This is independent of x. For $x > 3$, their utility is $4x - x^2 + \alpha\left(4x - \int_0^3 \frac{1}{3}y^2dy\right) = 6x - x^2 - \frac{1}{2}$. This is less than $\frac{17}{2}$, and therefore a regular altruist would not deviate from this distribution.

Material payoff resident regular altruist

The material payoff of the altruist, with $\alpha = \sigma = \frac{1}{2}$, is

$$\begin{aligned}
&\int_0^3 \left[\int_0^3 \frac{1}{3}\frac{1}{3} \cdot \left(4 \cdot \max(x,y) - x^2\right)\right] dydx \\
&= \frac{8}{9}\int_0^3 \left[\int_0^x x\right] dydx - \frac{1}{9}\int_0^3 \int_0^3 x^2 dydx \\
&= \frac{8}{9}\int_0^3 x^2 dx - \frac{1}{3}\int_0^3 x^2 dx = \frac{5}{9}\int_0^3 x^2 dx = \frac{5}{9}\left[\frac{1}{3}x^3\right]_0^3 = 5
\end{aligned}$$

Before going to a Homo Hamiltonensis, we compute the material payoff of a pure mutant.

Material payoff pure mutants

A mutant playing $y < 3$ with probability 1 earns a material payoff of

$$\begin{aligned}
&(1-\sigma)\left(\int_0^y \frac{4}{3}y\,dx + \int_y^3 \frac{4}{3}x\,dx\right) + \sigma 4y - y^2 \\
&= (1-\sigma)\left(\frac{4}{3}y^2 + \frac{2}{3}(3^2 - \cdot y^2)\right) + \sigma 4y - y^2 \\
&= (1-\sigma)\left(\frac{2}{3}y^2 + 6\right) + \sigma 4y - y^2
\end{aligned}$$

If we take $\sigma = \frac{1}{2}$, then this is

$$= \frac{1}{3}y^2 + 3 + 2y - y^2 = 3 + 2y - \frac{2}{3}y^2 < \frac{9}{2} < 5$$

Material payoff mutant Homo Hamiltonensis

A mutant Homo Moralis playing y would have expected utility

$$\begin{aligned}
&= (1-\kappa)\left((1-\sigma)\left(\int_0^y \frac{4}{3}y\,dx + \int_y^3 \frac{4}{3}x\,dx - y^2\right) + \sigma(4y - y^2)\right) + \kappa(4y - y^2) \\
&= (1-\kappa)(1-\sigma)\left(\frac{4}{3}y^2 + \frac{2}{3}(3^2 - \cdot y^2)\right) + ((1-\kappa)\sigma + \kappa)\,4y - y^2 \\
&= (1-\kappa)(1-\sigma)\left(\frac{2}{3}y^2 + 6\right) + ((1-\kappa)\sigma + \kappa)\,4y - y^2 \\
&= (1-\kappa)(1-\sigma)6 + ((1-\kappa)\sigma + \kappa)\,4y - \frac{3 - 2(1-\kappa)(1-\sigma)}{3}y^2
\end{aligned}$$

If we take $\kappa = \sigma = \frac{1}{2}$, then this is

$$= \frac{3}{2} + 3y - \frac{5}{6}y^2$$

which is maximized at $y = \frac{9}{5}$, where the material payoff of this mutant at $\epsilon = 0$ (as calculated just above) is below $\frac{9}{2}$, which is below 5.

C.4 Payoffs at $\epsilon > 0$ with a regular altruist as a resident

Equilibrium behaviour resident regular altruists for $\epsilon > 0$

Here we assume that the resident in the population is a regular altruist with $\alpha = \frac{1}{2}$, and that the mutant plays a pure strategy $0 \le z \le 3$. The best response against this pure strategy would be a uniform distribution on $\left[0, z - \frac{3}{4}\epsilon\right] \cup \left[z + \frac{3}{4}\epsilon, 3\right]$, which we denote by $G_{\epsilon,z}$; the utility that a

regular altruist derives from playing $x \in \left[0, z - \frac{3}{4}\epsilon\right] \cup \left[z + \frac{3}{4}\epsilon, 3\right]$ in this population is

$$\begin{aligned}
&(1-\epsilon+\epsilon\sigma)\int \left(\pi(x,y)+\alpha\pi(y,x)\right) dG_{\epsilon,z}(y) + (1-\sigma)\epsilon\left(\pi(x,z)+\alpha\pi(z,x)\right) \\
&= (1-\epsilon+\epsilon\sigma)\left((1+\alpha)\left(\int_0^x 4x\, dG_{\epsilon,z}(y) + \int_x^3 4y\, dG_{\epsilon,z}(y)\right) - x^2 - \alpha\left(\int y^2\, dG_{\epsilon,z}(y)\right)\right) \\
&\qquad + (1-\sigma)\epsilon\left((1+\alpha)4\max(x,z) - x^2 - \alpha z^2\right)
\end{aligned}$$

We can first consider $x \in \left[0, z - \frac{3}{4}\epsilon\right]$. In this case,

$$\int_0^x 4x\, dG_{\epsilon,z}(y) = \int_0^x \frac{1}{3(1-\epsilon/2)} \cdot 4x\, dy = \frac{4x^2}{3(1-\epsilon/2)}$$

and

$$\begin{aligned}
\int_x^3 4y\, dG_{\epsilon,z}(y) &= \int_x^{z-\frac{3}{4}\epsilon} \frac{1}{3(1-\epsilon/2)} \cdot 4y dy + \int_{z+\frac{3}{4}\epsilon}^3 \frac{1}{3(1-\epsilon/2)} \cdot 4y dy \\
&= \left[\frac{1}{3(1-\epsilon/2)} \cdot 2y^2\right]_x^{z-\frac{3}{4}\epsilon} + \left[\frac{1}{3(1-\epsilon/2)} \cdot 2y^2\right]_{z+\frac{3}{4}\epsilon}^3 \\
&= \frac{2}{3(1-\epsilon/2)}\left(9 - (z+\frac{3}{4}\epsilon)^2 + (z-\frac{3}{4}\epsilon)^2 - x^2\right) \\
&= \frac{2}{3(1-\epsilon/2)}\left(9 - 3\epsilon\cdot z - x^2\right)
\end{aligned}$$

With $\sigma = \alpha = \frac{1}{2}$, this makes the utility for $x \in \left[0, z - \frac{3}{4}\epsilon\right]$ equal to

$$\begin{aligned}
&(1-\frac{\epsilon}{2})\left(\frac{3}{2}\left(\frac{2}{3(1-\epsilon/2)}\left(9 - 3\epsilon\cdot z + x^2\right)\right) - x^2 - \frac{1}{2}\left(\int y^2\, dG_{\epsilon,z}(y)\right)\right) \\
&\qquad + \frac{\epsilon}{2}\left(\frac{3}{2}\cdot 4z - x^2 - \frac{1}{2}z^2\right) \\
&= 9 - \frac{2-\epsilon}{4}\left(\int y^2\, dG_{\epsilon,z}(y)\right) - \frac{\epsilon}{4}\cdot z^2
\end{aligned}$$

This is independent of x.

Then we can consider $x \in \left[z + \frac{3}{4}\epsilon, 3\right]$. In this case,

$$\begin{aligned}
\int_0^x 4x\, dG_{\epsilon,z}(y) &= \int_0^{z-\frac{3}{4}\epsilon} \frac{1}{3(1-\epsilon/2)} \cdot 4x\, dy + \int_{z+\frac{3}{4}\epsilon}^x \frac{1}{3(1-\epsilon/2)} \cdot 4x\, dy \\
&= \frac{4x}{3(1-\epsilon/2)}\left(x - (z+\frac{3}{4}\epsilon) + (z-\frac{3}{4}\epsilon) - 0\right) \\
&= \frac{4x}{3(1-\epsilon/2)}\left(x - \frac{3}{2}\epsilon\right)
\end{aligned}$$

and

$$\begin{aligned}\int_x^3 4y\, dG_{\epsilon,z}(y) &= \int_x^3 \frac{1}{3(1-\epsilon/2)} \cdot 4y dy \\ &= \left[\frac{1}{3(1-\epsilon/2)} \cdot 2y^2\right]_x^3 \\ &= \frac{2}{3(1-\epsilon/2)}\left(9-x^2\right)\end{aligned}$$

With $\sigma = \alpha = \frac{1}{2}$, this makes the utility for $x \in \left[z + \frac{3}{4}\epsilon, 3\right]$ equal to

$$\begin{aligned}&(1-\frac{\epsilon}{2})\left(\frac{3}{2}\left(\frac{2}{3(1-\epsilon/2)}\left(9-3\epsilon\cdot x+x^2\right)\right)-x^2-\frac{1}{2}\left(\int y^2\, dG_{\epsilon,z}(y)\right)\right) \\ &\qquad + \frac{\epsilon}{2}\left(\frac{3}{2}\cdot 4x - x^2 - \frac{1}{2}z^2\right) \\ &= 9 - \frac{2-\epsilon}{4}\left(\int y^2\, dG_{\epsilon,z}(y)\right) - \frac{\epsilon}{4}\cdot z^2\end{aligned}$$

This is also independent of x, and equal to the utility for $x \in \left[0, z - \frac{3}{4}\epsilon\right]$.

For values of x outside this interval, the utility is lower, and therefore a regular altruist would not deviate from the distribution of strategies.

Equilibrium behaviour mutant Homo Hamiltonensis for $\epsilon > 0$

A mutant Homo Moralis with $\kappa = \sigma = \frac{1}{2}$, playing z, with a resident regular altruist with $\alpha = \sigma = \frac{1}{2}$ best responding, would have expected utility

$$\begin{aligned}&(1-\kappa)\left((1-\sigma)(1-\epsilon)\int \pi(z,x) dG_{\epsilon,z}(x) + \left((1-\sigma)\epsilon+\sigma\right)\left(4z-z^2\right)\right) + \kappa(4z-z^2) \\ &= \frac{1}{2}\left(\frac{1-\epsilon}{2}\left(\int_0^z 4z\, dG_{\epsilon,z}(x) + \int_z^3 4x\, dG_{\epsilon,z}(x) - z^2\right) + \left(\frac{1+\epsilon}{2}\right)\left(4z-z^2\right)\right) + \frac{1}{2}(4z-z^2) \\ &= \frac{1}{2}\left(\frac{1-\epsilon}{2}\left(\int_0^{z-\frac{3}{4}\epsilon} \frac{1}{3(1-\epsilon/2)}\cdot 4z dx + \int_{z+\frac{3}{4}\epsilon}^3 \frac{1}{3(1-\epsilon/2)}\cdot 4x dx\right)\right) + \left(3+\epsilon\right)z - z^2 \\ &= \frac{1-\epsilon}{4}\left(\frac{4}{3(1-\epsilon/2)}\cdot z\left(z-\frac{3}{4}\epsilon\right) + \frac{2}{3(1-\epsilon/2)}\cdot\left(9-\left(z+\frac{3}{4}\epsilon\right)^2\right)\right) + \left(3+\epsilon\right)z - z^2 \\ &= \frac{1-\epsilon}{4}\left(\frac{2}{3(1-\epsilon/2)}\left(z^2 - 3\epsilon\cdot z + 9 - \frac{9}{16}\epsilon^2\right)\right) + \left(3+\epsilon\right)z - z^2 \\ &= \frac{1-\epsilon}{2(1-\epsilon/2)}\left(3-\frac{3}{16}\epsilon^2\right) + \left(3+\epsilon-\frac{1-\epsilon}{2(1-\epsilon/2)}\cdot\epsilon\right)z + \left(\frac{1-\epsilon}{6(1-\epsilon/2)} - 1\right)z^2 \\ &= \frac{1-\epsilon}{2(1-\epsilon/2)}\left(3-\frac{3}{16}\epsilon^2\right) + \frac{6-2\epsilon}{2-\epsilon}\cdot z - \frac{5-2\epsilon}{6-3\epsilon}\cdot z^2\end{aligned}$$

This is maximized at $z = \frac{3-\epsilon}{2-\epsilon}\frac{6-3\epsilon}{5-2\epsilon}$. For $\epsilon \downarrow 0$ this converges to $y = \frac{9}{5}$, which is the equilibrium value at $\epsilon = 0$.

Material payoff resident regular altruist for $\epsilon > 0$

We can compute the material payoff of a resident altruist θ playing this mixed BNE in a population with assortment parameter $\sigma = \frac{1}{2}$, and a mutant Homo Hamiltonensis τ that plays pure strategy $z = \frac{3-\epsilon}{2-\epsilon}\frac{6-3\epsilon}{5-2\epsilon}$, but we can also observe that, since $G_{\epsilon,z}$ converges almost surely to $G_{0,z}$, the material payoff will be continuous in ϵ, and converges to the payoff at 0 for $\epsilon \downarrow 0$

Material payoff mutant Homo Hamiltonensis for $\epsilon > 0$

We can compute the material payoff of a mutant Homo Hamiltonensis τ playing $z = \frac{3-\epsilon}{2-\epsilon}\frac{6-3\epsilon}{5-2\epsilon}$ in a population with assortment parameter $\sigma = \frac{1}{2}$, and a resident regular altruist θ that plays the mixed BNE, but we can also observe that, since $G_{\epsilon,z}$ converges almost surely to $G_{0,z}$, the material payoff will be continuous in ϵ, and converges to the payoff at 0 for $\epsilon \downarrow 0$

References

Akdeniz, A., and M. van Veelen (2020): "The cancellation effect at the group level," *Evolution*, 74(7), 1246–1254.

Alger, I., and J. W. Weibull (2012): "A generalization of Hamilton's rule—Love others how much?," *Journal of Theoretical Biology*, 299, 42–54.

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

——— (2016): "Evolution and Kantian morality," *Games and Economic Behavior*, 98, 56–67.

Bell, A. V., P. J. Richerson, and R. McElreath (2009): "Culture rather than genes provides greater scope for the evolution of large-scale human prosociality," *Proceedings of the National Academy of Sciences*, 106(42), 17671–17674.

Bergman, T. J., J. C. Beehner, D. L. Cheney, and R. M. Seyfarth (2003): "Hierarchical classification by rank and kinship in baboons," *Science*, 302(5648), 1234–1236.

Bomze, I., W. Schachinger, and J. Weibull (2020): "Does moral play equilibrate?," *Economic Theory*, pp. 1–11.

Call, J., and M. Tomasello (2008): "Does the chimpanzee have a theory of mind? 30 years later.," *Trends in Cognitive Sciences*, 12(5), 187.

Fischer, A., J. Pollack, O. Thalmann, B. Nickel, and S. Pääbo (2006): "Demographic history and genetic differentiation in apes," *Current Biology*, 16(11), 1133–1138.

Frank, R. H. (1987): "If homo economicus could choose his own utility function, would he want one with a conscience?," *The American Economic Review*, pp. 593–604.

——— (1988): *Passions within reason: The strategic role of the emotions*. WW Norton & Co.

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

——— (1964b): "The genetical evolution of social behaviour. II," *Journal of Theoretical Biology*, 7(1), 17–52.

Kant, I. (1785): *Grundlegung zur Metaphysik der Sitten*. [In English: Groundwork of the Metaphysics of Morals. 1964. New York: Harper Torch books.].

Krupenye, C., and J. Call (2019): "Theory of mind in animals: Current and future directions," *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(6), e1503.

Langergraber, K., G. Schubert, C. Rowney, R. Wrangham, Z. Zommers, and L. Vigilant (2011): "Genetic differentiation and the evolution of cooperation in chimpanzees and humans," *Proceedings of the Royal Society B: Biological Sciences*, 278(1717), 2546–2552.

Langergraber, K. E., J. C. Mitani, and L. Vigilant (2007): "The limited impact of kinship on cooperation in wild chimpanzees," *Proceedings of the National Academy of Sciences*, 104(19), 7786–7790.

Lieberman, D., J. Tooby, and L. Cosmides (2007): "The architecture of human kin detection," *Nature*, 445(7129), 727–731.

Miettinen, T., M. Kosfeld, E. Fehr, and J. Weibull (2020): "Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions," *Journal of Economic Behavior & Organization*, 173, 1–25.

Parr, L. A., and F. B. de Waal (1999): "Visual kin recognition in chimpanzees," *Nature*, 399(6737), 647–648.

Penn, D. C., K. J. Holyoak, and D. J. Povinelli (2008): "Darwin's mistake: Explaining the discontinuity between human and nonhuman minds," *Behavioral and Brain Sciences*, 31(2), 109–130.

Scally, A., B. Yngvadottir, Y. Xue, Q. Ayub, R. Durbin, and C. Tyler-Smith (2013): "A genome-wide survey of genetic variation in gorillas using reduced representation sequencing," *PLoS One*, 8(6).

Silk, J. B. (2009): "Nepotistic cooperation in non-human primate groups," *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533), 3243–3254.

Taylor, P. D. (1992a): "Altruism in viscous populations—an inclusive fitness model," *Evolutionary Ecology*, 6(4), 352–356.

——— (1992b): "Inclusive fitness in a homogeneous environment," in *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 249, pp. 299–302. The Royal Society.

van Veelen, M. (2006): "Why kin and group selection models may not be enough to explain human other-regarding behaviour," *Journal of Theoretical Biology*, 242(3), 790–797.

——— (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.

——— (2012): "Robustness against indirect invasions," *Games and Economic Behavior*, 74(1), 382–393.

——— (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.

van Veelen, M., and P. Spreij (2009): "Evolution in games with a continuous action space," *Economic Theory*, 39(3), 355–376.

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

Wilson, D. S., G. B. Pollock, and L. A. Dugatkin (1992): "Can altruism evolve in purely viscous populations?," *Evolutionary Ecology*, 6(4), 331–341.