The Endowment Effect as Blessing[*†]

Sivan Frenkel[‡] Yuval Heller[§] Roee Teper[¶]

June 15, 2017

Abstract

We study the idea that seemingly unrelated behavioral biases can coevolve if they jointly compensate for the errors that any one of them would give rise to in isolation. We suggest that the "endowment effect" and the "winner's curse" could have jointly survived natural selection together. We develop a new family of "hybrid-replicator" dynamics. Under such dynamics, biases survive in the population for a long period of time even if they only partially compensate for each other and despite the fact that the rational type's payoff is strictly larger than the payoffs of all other types.

Keywords: Endowment Effect, Winner's Curse, Bounded Rationality, Evolution.

JEL Classification: C73, D82, D03

1 Introduction

The growing field of Behavioral Economics has frequently identified differences between the canonical model of rational decision making and actual human behavior. These differences, usually referred to as "anomalies" or "biases," have been identified through controlled experiments in the laboratory, as well as in field experiments (see, e.g., Kagel & Roth, 1997;

[*] The manuscript was accepted for publication in the *International Economic Review* (**final pre-print**).

[†] A previous version of the paper was titled "Endowment as a Blessing." We thank Eddie Dekel, John Duffy, Alan Grafen, Richard Katzwer, Shawn McCoy, Erik Mohlin, Thomas Norman, Luca Rigotti, Larry Samuelson, Lise Vesterlund, the Associate Editor and the anonymous referees, and various seminar audiences for valuable discussions and suggestions, and to Sourav Bhattacharya for the query that initiated this project.

[‡] Coller School of Management, Tel Aviv University. frenkels@post.tau.ac.il. URL: https://sites.google.com/site/sivanfrenkel/.

[§] Department of Economics, Bar Ilan University. yuval.heller@biu.ac.il. URL: https://sites.google.com/site/yuval26/. The author is grateful to the European Research Council for its financial support (starting grant #677057).

[¶] Department of Economics, University of Pittsburgh. rteper@pitt.edu. URL: http://www.pitt.edu/∼rteper/.

Harrison & List, 2004). Such behavior is puzzling to economists, who are trained to think that competitive forces in our society and economy select optimal over suboptimal behavior. In this paper, we argue that sets of biases may persist because they jointly compensate for the errors that any one of them would give rise to in isolation. Thus, biases may coevolve as a "shortcut" solution that leads in specific important environments to behavior that is approximately optimal.[1] While the majority of the existing literature studies behavioral biases separately, our results suggest that one can gain a better understanding of different behavioral biases by analyzing their combined effects.

Our paper presents two key contributions. First, we show a relation between the endowment effect and the winner's curse that indicates that these biases approximately compensate for each other in barter trade and, therefore, could be the stable outcome of evolutionary dynamics. The *endowment effect* (Thaler, 1980) refers to an individual's tendency to place a higher value on a good once he owns it.[2] The *winner's curse,* or *cursedness* (Eyster & Rabin, 2005), is the failure of an agent to account for the informational content of other players' actions. Cursed agents underestimate the effect of adverse selection, and thus, for example, tend to overbid in common-value auctions and bilateral trade.[3] Though they are seemingly unrelated, we show that both biases may have coevolved together.

Our second key contribution is relevant to the study of relations between any pair of biases. In a seminal paper, Waldman (1994) shows that a pair of biases can be evolutionarily stable under sexual inheritance only if the level of each bias is optimal when taking the level of the other bias as fixed (Waldman calls such pairs "second-best adaptations").[4] Waldman (1994) studies a setup in which the set of levels of each bias is discrete and sparse, such that "second-best adaptations" exist. Our example of the winner's curse and the endowment effect, however, demonstrates that in various plausible setups the set of feasible biases is sufficiently dense and the payoff function is concave, and as a result there are no "second-best adaptations" except for having no biases at all. Therefore, applying Waldman's results to such setups does not deliver new predictions beyond that of a standard replicator-dynamics analysis.

[1] Cesarini *et al.* (2012) present experimental evidence suggesting that many common behavioral biases (and, in particular, loss aversion, which is closely related to the endowment effect) are partially heritable.

[2] See Kahneman *et al.* (1991) for a survey on the endowment effect, and Knetsch *et al.* (2001); Genesove & Mayer (2001); Bokhari & Geltner (2011); Apicella *et al.* (2014) for recent experimental evidence.

[3] See Kagel & Levin (2002, Chapter 1) for a survey on the winner's curse, and Grosskopf *et al.* (2007); Massey & Thaler (2013) for recent experimental support and field evidence. We follow Eyster & Rabin (2005) in the way we model the extent to which an agent is exposed to the winner's curse (see Section 2.2), and also in referring to this extent as "cursedness."

[4] For other methodological papers that study stable outcomes in sexual populations in which two traits co-evolve, see Karlin (1975); Eshel & Feldman (1984); Matessi & Di Pasquale (1996). Bergstrom & Bergstrom (1999) study the influence of sexual inheritance on parent-offspring conflicts.

This is where our analysis makes a contribution. We show that some pairs of biases can persist for relatively long periods of time even when there are no "second-best adaptations." Specifically, our results show that starting from any initial state, the population converges to a state in which agents have both biases and the level of each bias approximately compensates for the errors of the other bias. Following this convergence, there is a long process in which the population slowly drifts from having one pair of biases to having another pair in which the level of each bias is slightly lower. The length of this process (which eventually results in the population into having no biases) is inversely proportional to the extent to which each bias compensates for the other bias.[5]

Model. We demonstrate that the endowment effect and cursedness, while seemingly unrelated biases, compensate for each other in barter-trade interactions of the type that were common in prehistoric societies.[6] In our model, two traders each own a different kind of indivisible good and consider whether to participate in trade or not. The value of each good depends on an unobservable property that is known to the owner of the good but not to his trading partner. The potential gain of the traders also depends on additional conditions that are known to both players before they engage in trading, but can change from one instance of trade to another. Goods are exchanged if both traders agree to trade.

Each agent in the population is endowed with a pair of biases. The level of these biases determines his type, and the agent with a zero level of both biases (i.e., the unbiased one) is "rational." The first bias is *cursedness*, i.e., the extent to which an agent underestimates the relation between the partner's agreement to trade and the quality of this partner's good. In the trade game, agents in general choose to trade goods that are not very valuable and keep the valuable goods for themselves. A cursed trader does not pay enough attention to this fact and overestimates the value of his partner's good conditional on trade (in the extreme case, a fully cursed agent simply expects to get a good of ex-ante value). As a result, a more cursed agent will agree to trade goods with a greater personal value. Thus, cursedness leads agents to trade too much, and higher cursedness results in more trade.

The second bias is a *perception bias,* i.e., a function, ψ, that distorts an agent's subjective

[5] In addition, we make a technical contribution in extending Waldman's (1994) analysis from nonstrategic interactions, in which an agent's payoff depends only on his own decisions, to strategic interactions in which an agent's payoff depends also on the behavior of other agents in the population.

[6] Evidence from anthropology suggests that trade between groups, based on localization of natural resources and tribal specializations, was common in primitive societies (see Herskovits, 1952; Polanyi, 1957; Sahlins, 1972; Haviland *et al.*, 2007). Moreover, "[t]he literature on trade in nonliterate societies makes clear that barter is by far the most prevalent mode of exchange" (Herskovits, 1952, p. 188). Barter trade is not a central interaction in modern societies, but arguably, the time that has passed since the invention of currency is too short, in terms of genetic evolution, to eliminate attributes that were helpful for survival in prehistoric times.

valuation of his own good. If the good is worth x, an agent believes it to be worth $\psi(x)$. If $\psi(x) > x$, we say that the agent exhibits the endowment effect, but we allow agents to have other perception biases as well. An agent with the endowment effect does not agree to trade goods with low values since he believes those goods to be more valuable than they actually are and thus he loses profitable transactions.[7] Agents with the endowment effect trade too little, and traders with a higher level of endowment effect trade less.

Results. We analyze the interaction in a large population of traders with different levels of biases (types). Agents are randomly matched and play the barter game. We assume that agents do not observe the types of their partners and in each period they best-reply to the aggregate behavior. Their best reply, however, is distorted by their own biases. We show that there is a set Γ of types that exhibit both the winner's curse and the endowment effect, such that the two biases compensate each other. The set Γ includes not only the rational type, but also types at all levels of cursedness. Types who are more cursed exhibit a greater endowment effect.

Our first result (Proposition 1) shows that a distribution of types is a Nash equilibrium of the population game if and only if its support is a subset of Γ. Moreover, all agents in the population exhibit the same "as-if rational" behavior on the equilibrium path (their trading strategy is identical to that of a rational trader), and any type outside Γ achieves a strictly lower payoff if he invades the population. In a dynamic setting where the payoff of the barter game determine the agents' fitness and the frequency of types evolves according to a payoff-monotone selection (e.g., the replicator dynamics; see Taylor & Jonker, 1978), stable distributions of types are those with a support in Γ.

We then extend our analysis to the case where fitness is determined not only by the outcome of barter trade, but is also a result of other activities, in which the biases typically do not compensate for each other. We assume that while agents interact most of the time in barter trade, with a small probability p they play other games in which biases lead to strictly lower payoffs.[8] In this setup the rational type has a strict advantage, and as soon as a few

[7] The fact that traders may have an endowment effect seems at first sight at odds with experiments that have shown that professionals do not have an endowment effect for goods obtained solely for trade rather than personal consumption (e.g., Kahneman *et al.*, 1990; List, 2003, 2004; Lindsay, 2011). For example, a shoemaker will not have an endowment effect for the shoes he makes. This may be related to the fact that either the shoemaker does not have much value for the goods unless he can sell them or does not see them as his "endowment" in the first place because they are made for sale. However, manufacturing goods solely for trade is a feature of the modern world. Though tribes did specialize in specific goods for trade, traded goods were also consumed by other tribe members (Herskovits, 1952). Thus, there is no contradiction between the assumption that tribe members developed the endowment effect for their goods in primitive times and the empirical evidence that professional traders exhibit less of an endowment effect for traded goods.

[8] In Remark 2 we discuss how small values of p can also be interpreted as cases in which the other activities

rational "mutant" agents invade the population, the standard replicator dynamics converges to everyone being rational.[9]

We show, however, that for various plausible selection dynamics, the above result is no longer true. We present the family of *hybrid-replicator dynamics* in which, in contrast to the replicator dynamics, a newborn agent does not simply replicate the type of an incumbent. In such dynamics, an agent inherits with some probability each bias from a different incumbent, and with the remaining probability inherits both biases from a single incumbent. One plausible interpretation of this dynamics is a sexual inheritance where each offspring's genotype is a mixture of his parents' genes. Another interpretation is social learning in which some agents may learn different strategic aspects from different "mentors."

The hybrid-replicator dynamics is not payoff-monotone because only a fraction of the agent's offspring share his type, while the other offspring have "hybrid" types. Consider for example a population composed of a single type in Γ, that is, a type where the two biases compensate for each other. Now assume that this population is invaded by a small group of "mutant" rational agents. Such agents, by definition, have a higher fitness due to their advantage in non-barter activities. However, only a fraction of the rational agent's offspring are rational and this "dilutes" their relative fitness advantage. The remaining hybrid offspring have low fitness, because their single bias is not compensated by the other bias in the barter interaction. As a result, the biased incumbent is stable against unbiased mutants.

First, we use the new dynamic to extend Waldman's (1994) results to our setup and shows that only the rational type is stable against all types (Proposition 2). Next, we show that despite the former result, pairs of biases close to Γ can persist for long periods of time. Specifically, Proposition 3 shows a relatively quick global convergence to Γ: any type outside Γ can be eliminated by a "mutant" with the same cursedness level and a perception bias that is strictly closer to Γ. Moreover, a finite number of invasions by such mutants, which is independent of the initial state and of the frequency of the additional activities p, will bring the population very close to Γ.

Our last result (Proposition 4) shows that each type t in Γ is stable against all other

are frequent, but the negative net effect of each bias in these activities is small. In Section 6 we discuss how to adapt our results to a setup in which players interact each time in slightly different barter-trade interactions.

[9] In contrast to the assumption here that it is best to be rational, previous literature has suggested that in some setups each bias may have additional benefits. When biases are at least partially observable, the endowment effect can be beneficial by inducing credible toughness in bargaining (Heifetz & Segev, 2004; Huck *et al.*, 2005). The winner's curse can reduce the risk of creating information cascades e.g., Bernardo & Welch (2001), show how an evolutionary process can induce agents to underestimate information that is revealed by the actions of others. In addition, developing fully rational thinking may incur "complexity costs" that are abstracted away in the model. See also Compte & Postlewaite (2004); Robson & Samuelson (2009) for other examples of the benefits of biases.

types, except for a "mutant" type that is very close to t but has slightly lower levels of each bias. This implies that in a small neighborhood around Γ, the population slowly drifts toward the rational type. Each step in this drift requires the appearance of a new mutant with slightly smaller biases, and the length of the sequence of invasions that eventually takes the population all the way to the rational type is at least $O\left(\frac{1}{p}\right)$. Thus, for low levels of p, the population eventually reaches the rational type only after a very long time.[10]

Related Literature and Contribution. Our paper is related to the "indirect evolutionary approach" literature (Guth & Yaari, 1992), which deals with the evolution of preferences that deviate from payoff (or fitness) maximization.[11] A main stylized result in this literature (see Ok & Vega-Redondo, 2001; Dekel *et al.*, 2007) is that biases can be stable only if types are observable, and so a player can condition his behavior on an opponent's type.[12] By contrast, we show that even with the "conventional" replicator dynamics, stable states may contain biased players who play as if they were rational on the equilibrium path (off the equilibrium path, however, their "mistakes" can be observed).[13] Moreover, when considering hybrid-replicator dynamics, players can also play suboptimally on the equilibrium path.

Our paper is also related to the literature that explains how behavioral biases may evolve. A majority of these papers deal with a single bias. A few papers have dealt with the possibility that evolution creates two biases that are significantly different and yet complementary. Heifetz *et al.* (2007) develop a general framework in which natural selection may lead to perception biases, and show that if preferences are observable, then, generically, non-material preferences are stable. In a non-evolutionary context, Kahneman & Lovallo (1993) suggest that two biases, namely, excessive risk aversion and the tendency of individuals to consider decision problems one at a time, partially cancel each other out. Recently, Ely (2011) demonstrated that in evolutionary processes improvements tend to come in the form of "kludges," that is, marginal adaptations that compensate for, but do not eliminate, fundamental design inefficiencies. Johnson & Fowler (2011) show that overconfidence arises naturally in a setup

[10] Studies of biological evolutionary processes have discussed population states that are stable only in limited horizons but not in the "long run," and have demonstrated that human populations can be found today in such "quasi-stable" states, even after many thousands of years of evolution (see, e.g., Hammerstein, 1996).One famous example of this in the human population is sickle cell disease, which occurs when a person has two mutated alleles. This disease is relatively frequent, especially in areas in which malaria is common, due to the heterozygote advantage: a person with a single mutated allele has a better resistance to malaria.

[11] See Remark 1 below for a discussion on extending this literature to dealing with biases.

[12] A related example is Huck *et al.* (2005) and Heifetz & Segev (2004) who show that an endowment effect observed by others can evolve in populations that engage in bargaining, through its use as a "commitment" device. Herold (2012) assumes serviceability when analyzing the co-evolution of preferences for punishing and preferences for rewarding. See also Heller (2015) for a related result arising from limited foresight.

[13] See Sandholm (2001) for a related result in a setup of preference evolution.

where agents are not fitness-maximizing and use a non-Bayesian decision-making heuristic. Bénabou & Tirole (2002) show that overconfidence may be optimal when agents have time-inconsistent preferences. Finally, Herold & Netzer (2015) show that, different biases that are postulated in prospect theory partially compensate for each other, and Steiner & Stewart (2016) show that noise in information processing may be mitigated by over-weighting of small probability events. To the best of our knowledge, the present paper is the first to tie together the winner's curse and the endowment effect.

The paper is organized as follows. Section 2 presents the basic model, which is analyzed in Section 3. In Section 4 we introduce additional activities and the hybrid-replicator dynamics. Section 5 presents the main results. We discuss some of our assumptions in Section 6. Section 7 briefly concludes. All proofs appear in the Appendix.

2 Basic Model

We present a model of barter trade, in which a population consists of a continuum of agents who are randomly matched to engage in a trading interaction. Each agent in the population is endowed with a type that determines his biases. Agents do not observe the types of their trading partners. We describe below the different components of the interaction between each pair, and then proceed to describe the induced population game.

2.1 Barter Interaction

A barter interaction is composed of two agents matched as trading partners. Each agent $i \in \{1, 2\}$ in the pair owns a different kind of indivisible good, and observes privately the value of his own good, $\mathbf{x}_i$. We assume $\mathbf{x}_1, \mathbf{x}_2$ are continuous, independent, and identically distributed with full support over $[L, H]$, where $0 < L < H$. Let $\mu \equiv E(\mathbf{x}_i)$ be the (ex-ante) expected value of $\mathbf{x}_i$ and let $\mu_{\leq y} \equiv E(\mathbf{x}_i | \mathbf{x}_i \leq y)$ be the expected value of $\mathbf{x}_i$ given that its value is at most y.

Both traders receive a public signal $\alpha \geq 1$, which is a "surplus coefficient" of trade: the good owned by agent $-i$ is worth $\alpha \cdot \mathbf{x}_{-i}$ to agent i. High α represents better conditions for trade independently of the quality of the goods. For example, if both parties have a great need for the commodity they do not own, then α are high. Given that α denotes trade conditions other than quality, which is represented by $\mathbf{x}_1$ and $\mathbf{x}_2$, we assume that α, $\mathbf{x}_1$, and $\mathbf{x}_2$ are all independent. The coefficient α can have any continuous distribution with support[14]

[14] Full support over $\left[1, \frac{H}{L}\right]$ implies that for each fixed cursedness level, any two different perception biases induce different threshold strategies. Without this property our results would not change qualitatively: the

$\left[1, \frac{H}{L}\right]$. The agents interact by simultaneously declaring whether they are willing to trade. The goods are exchanged if and only if both agents agree to trade.

2.2 Biases / Types

Each agent has a pair of biases, and their specific levels are denoted by the agent's type, $t = (\chi, \psi)$. The first bias is *cursedness* à la Eyster & Rabin (2005). A trader of type $\chi \in [0,1]$ best-replies to a biased belief that the expected value of his partner's good is $\alpha \cdot (\chi \cdot \mu + (1-\chi) \cdot \mu_\alpha)$, where μ_α is the expected value of his partner's good when the partner agrees to trade and the trade coefficient is α. Thus, a cursed trader only partially takes into account the informational content of the other trader's action (a rational agent has $\chi = 0$; a "fully cursed" trader with $\chi = 1$ believes he always gets an "average" object). Notice that if $\mu_\alpha < \mu$ (as we show below), then a χ-cursed trader with $\chi > 0$ overestimates the quality of his partner's good.[15]

The second component, ψ, is a trader's *perception bias* regarding his own good. We assume that $\psi \in \Psi$, where Ψ is the set of continuous and strictly increasing functions from $[L, H]$ to itself. A trader with perception bias ψ best-replies to a biased belief that his own good's value is $\psi(x)$ (rather than x). If $\psi(x) > x$ for all $x \neq H$ we say that the trader exhibits an *endowment effect*. Given two perception biases, ψ_1 and ψ_2, we say that ψ_1 is *more biased* than ψ_2, denoted by $\psi_1 \succeq \psi_2$, if for all $x \in [L, H]$ either $\psi_1(x) \leq \psi_2(x) \leq x$ or $\psi_1(x) \geq \psi_2(x) \geq x$. We write $\psi_1 \succ \psi_2$ if $\psi_1 \succeq \psi_2$ and $\psi_2 \not\succeq \psi_1$. Letting $I \in \Psi$ denote the identity function, $I(x) \equiv x$, type $(0, I)$ is the unbiased (or "rational") type. Denote by $T \equiv [0,1] \times \Psi$ the set of all types.

2.3 Strategies and Configurations

A general strategy for trader i is a function from α to values of $\mathbf{x}_i$ for which the trader agrees to trade. If an agent expects a positive surplus from trading an object of value x_i, then he expects a positive surplus also from trading objects with a value less than x_i. Thus, we can restrict our attention to threshold strategies. An agent's *pure threshold strategy* (or

set Γ we define below would include more types, but the observed behavior of these types would be the same. Finally, the results would be similar if we allowed for smaller or greater α's, or if we assume full support only over $\left[1, \frac{H}{\mu}\right]$; however, this would make the notation cumbersome.

[15] The effect here is similar to the "winner's curse" in common-value auctions, where cursed agents overbid because they do not fully take into account the fact that, conditional on winning the auction, other bidders have lower signals, and thus the expected value of the good is lower than their own signal suggests. This behavior is observed in experiments (Kagel & Levin, 2002, Chapter 1). See Eyster & Rabin (2005) for further discussion.

simply, strategy) is a continuous function $s : \left[1, \frac{H}{L}\right] \to [L, H]$ that determines, for each α, the maximal value of x for which the agent accepts trade.[16] That is, an agent who follows strategy s accepts trade if and only if $x \leq s(\alpha)$. Note that $\mathbf{x}_i$ is continuous and so there is always a unique best response (see Equation 1 below), and hence the focus on pure strategies is without loss of generality. Let S denote the set of strategies. We say that a strategy s is strictly increasing if $\alpha > \alpha'$ implies that $s(\alpha) \geq s(\alpha')$ with equality only if $s(\alpha') = H$.

In what follows we assume that the number of types in the population is finite. Specifically, let $\Delta(T)$ be the set of type distributions with finite support (we slightly abuse notation and denote by t degenerate distributions with a single type t). A state of the population, or "configuration," is formally defined as follows.

Definition 1. A *configuration* is a pair (η, b), where $\eta \in \Delta(T)$ is a distribution of types and $b : \text{supp}(\eta) \to S$ is a *behavior* function assigning a strategy to each type.

Observe that the definition implies that an agent does not observe his opponent's type. He best-replies while taking into account the value of a random traded good, which is determined jointly by the distribution of types (biases) in the population and the strategy that each type uses in the game.[17] Given a configuration (η, b), let $\mu_\alpha(\eta, b)$ be the mean value of a good of a random trader, conditional on the trader's agreement to trade when the trade coefficient is α (formally defined in the Appendix in (5)). Notice that this mean value is determined jointly by the distribution of types (biases) in the population and the strategy each type employs in the game. Let $s_t^*(\alpha)(\eta, b)$ denote the best-reply threshold of a trader of type $t = (\chi, \psi)$ who is facing a surplus coefficient α and configuration (η, b). Formally, when $\alpha \cdot (\chi \cdot \mu + (1-\chi) \cdot \mu_\alpha(\eta, b)) \leq \psi(H)$, then $s_t^*(\alpha)(\eta, b)$ is the unique value in $[L, H]$ that solves the equation

$$\psi\left(s_t^*(\alpha)(\eta, b)\right) = \alpha \cdot \left(\chi \cdot \mu + (1-\chi) \cdot \mu_\alpha(\eta, b)\right). \tag{1}$$

The interpretation of (1) is as follows. The RHS describes the value of the good a trader expects to receive from a trade, given his cursedness level χ. The LHS describes the value the trader attaches to a good of value $s_t^*(\alpha)(\eta, b)$, given his perception bias ψ. A trader strictly prefers to trade if and only if his good's perceived value is less than the expected value of his partner's good, conditional on this partner agreeing to trade. If the value that the trader expects to receive from trade is lower than H, then $s_t^*(\alpha)(\eta, b)$ is the unique threshold for which the trader is indifferent between trading and not trading. When $\alpha \cdot$

[16] Without the mild assumption of continuity in α there may be some Nash equilibria that differ from elements of Γ only with respect to values of x that do not serve as a threshold for any type.

[17] In Remark 1 below we discuss how such a best reply may be the result of a simple learning process.

$(\chi \cdot \mu + (1-\chi) \cdot \mu_\alpha(\eta, b)) > \psi(H)$, that is, the expected value of a good obtained by trade is strictly higher than the perceived value of the owner's good, then $s_t^*(\alpha)(\eta, b) \equiv H$ (recall that by definition $\psi(H) \leq H$).

We conclude by describing the influence of biases on behavior, as implied by (1):

1. A cursed trader ($\chi > 0$) overestimates the good of his partner and therefore uses a threshold that is too *high* (since we deal with threshold strategies $\mu_\alpha \leq \mu$). That is, cursed traders trade too much.

2. A trader with an endowment effect ($\psi(x) > x$) overestimates the quality of his own good and therefore uses a threshold that is too *low* and trades too little.

2.4 Equilibrium Configurations and the Population Game

A configuration is an equilibrium if each type best-replies in the manner presented above.

Definition 2 (Equilibrium Configuration)**.** A configuration (η, b) is an equilibrium if for each type $t = (\chi, \psi) \in \text{supp}(\eta)$ and for every $\alpha \in \left[1, \frac{H}{L}\right]$, $b(t)(\alpha) = s_t^*(\alpha)(\eta, b)$.

Our first result, formally presented below, shows the existence of equilibrium configurations, and that in all equilibrium configurations agents trade more when the surplus coefficient α is greater. However, there may be more than one equilibrium configuration with the same underlying distribution. In what follows, therefore, we assume that the equilibrium configuration is chosen according the continuous function b^*, i.e., that each population η plays the equilibrium configuration $\left(\eta, b_\eta^*\right)$. The lemma below also shows that such a *continuous* function b^* exists.

Lemma 1 (Existence and Selection of Equilibrium Configurations)**.** *For every distribution of types $\eta \in \Delta(T)$, there exists a behavior b such that (η, b) is an equilibrium configuration. Moreover, (1) in any such equilibrium configuration (η, b), the strategy $b(t)$ is strictly increasing for any type $t \in supp(\eta)$, and (2) there exists a continuous selection function b_η^* that assigns to each type distribution $\eta \in \Delta(T)$ a behavior such that $\left(\eta, b_\eta^*\right)$ is an equilibrium configuration.*

One example of b^* is a function that chooses the equilibrium with maximal trade; but any continuous function will do. The continuity of the equilibrium selection function b^* is required for the payoff function in the population game (Defined below) to be continuous. Since we assume that for each distribution η a specific equilibrium is played, in what follows we can focus solely on the distribution of types η.

The barter-trade interaction together with the equilibrium selection function b^* define a population game $G_0 = (T, u)$, where T is the set of types (as defined above), and $u : T \times \Delta(T) \to [L, H]$ is a continuous payoff function that describes the expected value of a good obtained by a type-t agent who best-replies to a trade with a random partner from configuration $\left(\eta, b^*_\eta\right)$; a formal definition of $u(t, \eta)$ appears in Appendix A.2.

Remark 1 (Learning to Best-Reply). Our notion of population game applies the "indirect evolutionary approach" of a two-layer evolutionary process: a slower process according to which the distribution of types evolves, and a faster process according to which agents learn to "subjectively" best-reply to the aggregate behavior in the population. Most of the existing literature studies a setup in which types represent subjective preferences (see, e.g., Guth & Yaari, 1992; Ok & Vega-Redondo, 2001; Dekel *et al.*, 2007). It is reasonable to ask whether the assumption of subjective best-replying is still appropriate when dealing with behavioral biases such as cursedness, where agents make "mistakes."[18] That is, one may criticize the implicit assumption that agents perfectly understand the aggregate biased behavior of other agents, but then resort to a biased best reply. Observe, however, that the only information an agent requires for best-replying is an estimator of the expected value of a traded good. Consider agents who know the unconditional expected value of each good, and observe the value of traded goods in several past interactions. Given our definitions, non-cursed agents use the average value of previously traded goods as an estimate of the value of traded goods, while fully-cursed agents ignore past observations and believe that traded goods have, on average, the same value as non-traded goods.[19] Partially-cursed agents will pay only partial attention to historical values. Thus, the behavior we describe as a biased self-reply may arise from plausible "mistakes" in the course of a simple learning process.

3 Compensation of Biases in Barter Trade

In this section we analyze the equilibrium of the population game described above.

[18] We focus on cursedness since the endowment effect has a natural interpretation as being part of subjective preferences.

[19] Idiosyncratic errors due to finite samples do not have any qualitative effect on our results, as they induce a similar outcome to the random traits discussed in Section 6.

3.1 Rational and As-If Rational Behavior

As a first step in analyzing G_0, consider the case where the entire population is unbiased; that is, the threshold of each agent is determined by the indifference condition

$$x^*(\alpha) \equiv s^*_{(I,0)}(\alpha)(I,0) = \min\left\{\alpha \cdot \mu_\alpha(I,0), H\right\},$$

which is derived by substituting $\psi(x) = x$ and $\chi = 0$ into (1), and using the minimum to ensure that $x^*(\alpha) \leq H$. It is easy to show that if all agents play homogeneously in such a way, then in a Nash equilibrium the "rational" threshold, denoted by x^*, must be a solution to the equation

$$x^*(\alpha) = \min\left\{\alpha \cdot \mu_{<x^*(\alpha)}, H\right\}.$$

Next, associate with each $\chi \in [0,1]$ the perception bias $\psi^*_\chi \in \Psi$ defined by

$$\psi^*_\chi(x) \equiv \chi \cdot \frac{\mu}{\mu_{\leq x}} \cdot x + (1-\chi) \cdot x. \tag{2}$$

Now let $\Gamma = \left\{\left(\chi, \psi^*_\chi\right) : \chi \in [0,1]\right\} \subset T$ be the set of all such types. Observe that: (1) $\psi^*_0(x) \equiv I$ (thus the rational type $(0, I)$ is in Γ); (2) all other types in Γ are cursed ($\chi > 0$) and exhibit the endowment effect ($\psi^*_\chi(x) > x$ for all $x < H$); and (3) types in Γ who are more cursed also have a larger endowment effect: $\chi_1 > \chi_2$ implies that $\psi^*_{\chi_1} \succ \psi^*_{\chi_2}$.

Assume that the population behaves according to the "rational Nash equilibrium" described above, and therefore $\mu_\alpha = \mu_{<x^*(\alpha)}$. Then, the threshold chosen by types in Γ is defined by the equality

$$\chi \cdot \frac{\mu}{\mu_{\leq s^*_t(\alpha)}} \cdot s^*_t(\alpha) + (1-\chi) \cdot s^*_t(\alpha) = \alpha \cdot \left(\chi \cdot \mu + (1-\chi) \cdot \mu_{<x^*(\alpha)}\right)$$

in the case where$\alpha \cdot (\chi \cdot \mu + (1-\chi) \cdot \mu_\alpha) \leq H$, and $s^*_t(\alpha) = H$ otherwise. That is,

$$s^*_t(\alpha) = \min\left\{\alpha \cdot \mu_{<x^*(\alpha)}, H\right\} = x^*(\alpha).$$

In words, when all other agents in the population behave "rationally," the behavior of types in Γ is indistinguishable from the rational type. The endowment effect and cursedness compensate for each other, and the $\left(\chi, \psi^*_\chi\right)$ agent behaves (on the equilibrium path) as if he were rational.

3.2 Equilibrium and Stability

We now formalize the ideas that were presented informally above. First, let us define a Nash equilibrium of the population game, as a distribution of types that is a best reply to itself.

Definition 3. Distribution $\eta \in \Delta(T)$ is a *Nash equilibrium* in game $G_0 = (T, u)$ if $u(t, \eta) \geq u(t', \eta)$ for all $t \in supp(\eta)$ and $t' \in T$. It is a *strict Nash equilibrium* if the inequality is strict for each $t \neq t'$.

The following definition ensures that a set of types has two properties: (1) the set is internally stable since all types have the same payoff (internal equivalence) and (2) the set is immune to an invasion by a small number of "mutant" types, since those types are outperformed by the types in the set (external strictness).

Definition 4. A set of types $Y \subseteq T$ is *internally equivalent* and *externally strict* in game $G_0 = (T, u)$ if $u(t, \eta) = u(t', \eta) > u\left(\hat{t}, \eta\right)$ for all $\eta \in \Delta(Y)$, $t, t' \in Y$, and $\hat{t} \in T \backslash Y$.

Using these definitions, we can explicitly formulate our first main result:

Proposition 1. *A distribution $\eta \in \Delta(T)$ is a Nash equilibrium of G_0 if and only if $supp(\eta) \subseteq \Gamma$. Moreover, Γ is an internally equivalent and externally strict set.*

In Appendix A.4 we show that Proposition 1 implies that the set Γ is *asymptotically stable in payoff monotone selection dynamics,* such as the replicator dynamics (Taylor & Jonker, 1978).

4 Additional Activities

In this section we first introduce into our basic model additional activities in which biases have negative impacts, present the "hybrid-replicator dynamics," and motivate the focus on homogeneous populations.

4.1 Introducing Additional Activities

The main drawback of the analysis in the previous section is the assumption that fitness is the result of a single interaction. Obviously, agents engage in many activities that determine their fitness, and it is reasonable to assume that in some of these activities biases will have a negative impact on the biased agent's payoff. For example, in a trade of goods with a publicly observable quality, or goods with a private value, cursedness is not relevant, and the endowment effect results in an average loss. In what follows we allow agents to engage in additional activities, and assume that in such activities biases are harmful. In this section, moreover, we show that in this case the rational type has an advantage over all the other types, including those in Γ. In the next section we show that with some plausible assumptions on the dynamics, biased types can be stable.

We model the additional activities as follows. With some probability p agents take part in other activities in which biases are detrimental. Formally, for $0 \leq p \leq 1$, agents play a population game $G_p = (T, u_p)$ with a payoff function $u_p : T \times \Delta(T) \to [L, H]$ defined as

$$u_p(t, \eta) = (1 - p) \cdot u(t, \eta) + p \cdot v(t, \eta), \tag{3}$$

where $v(t, \eta)$, the payoff in other activities, is *bias-monotone:* larger biases yield lower payoffs. Formally, for every $(\chi, \psi), (\chi', \psi') \in T$, and $\eta \in \Delta(T)$:

$$\chi < \chi' \text{ and } \psi \prec \psi' \Rightarrow v((\chi', \psi'), \eta) < v((\chi, \psi), \eta).$$

In addition, we assume that $v(t, \eta)$ is *Lipschitz-continuous.*

Remark 2 (Interpreting Small Values of p). Most of our results assume that p is sufficiently small (i.e., $0 < p \ll 1$). In addition to capturing infrequent additional activities, one can also reinterpret these small values of p as capturing frequent additional activities in which each bias is (on average) only slightly detrimental. Consider a setup in which there are many possible non-barter trade interactions. In some of these interactions the biases are payoff-irrelevant, in others they are beneficial (see the examples discussed in Footnote 9), while in others they are detrimental. We assume that the net effect of each bias in these additional activities is negative, but small. For concreteness, assume that the probability of having a non-barter activity is q and that the order of magnitude of the negative net effect of each bias is ϵ. In this setup, one should interpret $p = q \cdot \epsilon$ as describing both the frequency of the additional activities and the net effect of each bias in these activities. In particular, small values of p may correspond to small negative net effects (i.e, $\epsilon \ll 1$), rather than to small frequencies of non-barter activities. Thus, our assumption that p is small reflects the stylized fact that barter trade was a central interaction in hunter-gatherer societies (as discussed in Footnote 6), and that the biases had small negative net effects in the non-barter activities.

By definition, in any game G_p the rational type $(0, I)$ has maximal v. In the previous section we showed that in G_0 there are types that behave as if they were rational in some configurations, but clearly this is not the case for any $p > 0$. In such cases the rational type always has an advantage. Therefore, under payoff-monotone dynamics, regardless of the initial state of the population, as soon as the rational type $(0, I)$ invades the population he strictly outperforms all incumbents, and takes over the entire population. Payoff-monotone dynamics, however, are not the only reasonable type of dynamics to assume. In the following subsection we introduce a plausible family of non-payoff-monotone dynamics that includes, among other things, biological heredity.

4.2 Hybrid-Replicator Dynamics

We model the evolutionary selection process in discrete time with each period τ representing a generation. The dynamics is represented by a deterministic transition function $g : \Delta(T) \to \Delta(T)$ describing the distribution of types in the next generation as a function of the distribution in this generation. The family of hybrid-replicator dynamics is characterized by a *recombination rate* $r \in [0,1]$ that describes the probability that each offspring is randomly assigned to two incumbent agents ("parents") and copies a single trait (in our case, a bias) from each one of them; under the complementary probability each offspring is assigned to a single incumbent and copies both of its traits. As in the replicator dynamics, these random assignments are distributed according to the incumbents' fitness. Below we provide a semi-formal description of the transition function. Formal definitions appear in Appendix A.5.

As is common in the literature, we assume that the expected number of offspring of each agent is equal to his game payoff plus a positive constant that reflects background factors that are unrelated to the game. Let $f_\eta(t) \in R^+$ be the relative fitness of type t in population $\eta \in \Delta(T)$, that is, this agent's fitness divided by the average fitness in the population. Denote by $\eta(\chi)$ and $\eta(\psi)$ the total frequency of types in η with cursedness level $\chi \in [0,1]$ and perception bias $\psi \in \Psi$, respectively. Now let $f_\eta(\chi)$ and $f_\eta(\psi)$ be the expected relative fitness of types in η with cursedness level χ and perception bias ψ, respectively. Thus, $\eta(t) \cdot f_\eta(t)$ is the probability of drawing type t from a population η after "reproduction." Similarly, $\eta(\chi) \cdot f_\eta(\chi)$ and $\eta(\psi) \cdot f_\eta(\psi)$ are the probabilities of drawing an agent with cursedness level $\chi \in [0,1]$ and perception bias $\psi \in \Psi$, respectively, after reproduction. Thus, the transition function is defined as follows:

$$g(\eta)((\chi,\psi)) = (1-r) \cdot \eta((\chi,\psi)) \cdot f_\eta((\chi,\psi)) + r \cdot \eta(\chi) \cdot f_\eta(\chi) \cdot \eta(\psi) \cdot f_\eta(\psi). \tag{4}$$

The family of hybrid-replicator dynamics extends the standard replicator dynamics (Taylor & Jonker, 1978; reformulated in Weibull, 1997, Chapter 4.1) for which $g(\eta)((\chi,\psi)) = \eta((\chi,\psi)) \cdot f_\eta((\chi,\psi))$. Observe that a hybrid-replicator dynamics coincides with the replicator dynamics if either $r = 0$ or one of the biases has the same level in the entire population.[20]

One interpretation of the hybrid-replicator dynamics is biological heredity. If one assumes that each trait (i.e., bias) is controlled by a different locus (i.e., position in the DNA sequence), then the probability that each child inherits each trait from a different parent is equal to the biological recombination rate between these loci. This parameter is equal to 0.5 if the two loci are in two different chromosomes, and it is strictly between 0 and 0.5 if the two loci are in different locations in the same chromosome. A hybrid-replicator dynamics is an exact description of a selection process in a haploid population in which each individual carries one copy of each chromosome and each trait is controlled by a single locus. At the same time, it is a stylized description that captures the important relevant properties of a selection process in a diploid population, like the human population, in which each individual carries two copies of each chromosome and each trait is controlled by several loci (see, e.g., Maynard Smith 1971; Liberman 1988). One can show that the hybrid-replicator dynamics has the same implications for asymptotic stability as a more detailed description of diploid populations.[21]

An additional interpretation is that of a social learning process. The parameter r determines the frequency of new agents who independently choose two "mentors" and imitate a single trait of each; each of the remaining new agents chooses a single "mentor" and imitates both his traits.

4.3 Instability of Heterogeneous Populations

As is common in the literature (see, e.g., Waldman, 1994; Alger & Weibull, 2013), our formal results in Section 5 focus on studying the stability of homogeneous populations, which include a single incumbent type, against an invasion by a single mutant type, rather than studying the stability of heterogeneous distributions against an invasion by heterogeneous groups of mutant types. The restriction to a single mutant type reflects the assumption that mutants are rare; it can be relaxed without affecting the results (but it demands more cumbersome notation and definitions). Thus, in this section where we informally sketch why heterogeneous populations are not stable in our setup, the restriction to a single incumbent type is motivated.

We begin by demonstrating why a heterogeneous distribution η that includes two types in Γ (t and t') is unstable. Observe that the support of a fixed point of a hybrid-replicator

[20] If all types have the same level of cursedness, χ, then $\eta(\chi) = 1$, $\eta(\psi) = \eta((\chi, \psi))$, $f_\eta(\chi) = 1$, and $f_\eta(\psi) = f_\eta((\chi, \psi))$, which implies that this dynamics coincides with the replicator dynamics. The same happens when there is a single ψ.

[21] In particular, assuming that each trait is influenced by a different chromosome and that the "mutant" type can be "dominant" in both loci yields $r = 0.5$. Other plausible assumptions (e.g., "recessive" mutants, and the two loci being in different parts of the same chromosome) yield a value strictly between 0 and 0.5.

dynamics (with $r > 0$) must be a product set (i.e., $\mathcal{X}_\eta \times \Psi_\eta$), and thus η also includes the two "hybrid types" that combine a trait from each of these types:

$$\left\{t = \left(\chi, \psi^*_{\chi}\right), t' = \left(\chi', \psi^*_{\chi'}\right), \left(\chi, \psi^*_{\chi'}\right), \left(\chi', \psi^*_{\chi}\right)\right\}.$$

For simplicity, assume that types t and t' have the same weight in the population, $\eta\left(t\right) = \eta\left(t'\right)$, and thus also $\eta\left(\left(\chi, \psi^*_{\chi'}\right)\right) = \eta\left(\left(\chi', \psi^*_{\chi}\right)\right)$. Denote the average cursedness by $\bar{\chi} = 0.5 \cdot (\chi + \chi')$. When p is low, the distribution η is not stable against an external mutant with type $\bar{t} = \left(\bar{\chi}, \psi^*_{\bar{\chi}}\right)$. This "average" type outperforms the incumbents because: (1) $\bar{t}$ has approximately the same fitness as types t and t'; and (2) the "hybrid siblings" of $\bar{t}$ that have only one of his traits, $\left(\bar{\chi}, \psi^*_{\chi}\right)$, $\left(\bar{\chi}, \psi^*_{\chi'}\right)$,$\left(\chi, \psi^*_{\bar{\chi}}\right)$, and $\left(\chi', \psi^*_{\bar{\chi}}\right)$, are substantially closer to Γ and, therefore, perform better than the hybrid incumbent types $\left(\chi, \psi^*_{\chi'}\right)$ and $\left(\chi', \psi^*_{\chi}\right)$. (This is because $|\chi - \bar{\chi}|$ and $|\chi' - \bar{\chi}|$ are less than $|\chi - \chi'|$.) Thus, on average, $\bar{t}$'s "siblings" achieve a higher payoff in the barter trade. This implies that $\bar{\chi}$ and $\psi^*_{\bar{\chi}}$ have a higher average fitness than the other biases (i.e., $f_\eta\left(\bar{\chi}\right) > f_\eta\left(\chi\right), f_\eta\left(\chi'\right)$ and $f_\eta\left(\bar{\psi}\right) > f_\eta\left(\psi\right), f_\eta\left(\psi'\right)$). By (4), a hybrid-replicator dynamics implies that mutant $\bar{t}$ succeeds in invading the population (his fitness is as high as that of the incumbents, and the fitness of each of his traits is strictly higher than that of the incumbents' traits).

This argument can be extended to general heterogeneous distributions. Observe that the payoff of the barter trade is strictly concave in a trader's thresholds (see part 2 in the proof of Proposition 1, and assume that trade occurs with positive probability). In addition, the thresholds of the traders are strictly increasing (decreasing) in the cursedness level (perception bias). These two observations imply that "intermediate" types use intermediate thresholds; that is, if each bias of type t is a mixed average of the respective biases of types t_1 and t_2, then its threshold strategy is strictly between the threshold strategies of types t_1 and t_2. If a heterogeneous population, η, is a fixed point of the dynamics, then different types must use different thresholds (because the support of the population is a product set, and two types that differ in only one of the traits must have different threshold strategies). Intuitively, due to these observations, there is a "mean" type $\bar{t}$ with biases that are weighted averages of the incumbents' biases, such that it uses thresholds that are weighted averages of the incumbents' thresholds, and a similar property holds for his "hybrid" offspring. (The explicit expression of $\bar{t}$ involves technical difficulties and this is why we are only sketching an intuitive argument.) Finally, due to strict concavity, mutant agents with type $\bar{t}$ and their "hybrid" offspring achieve, on average, strictly higher payoffs in the barter trade relative to the incumbents. For a sufficiently low p, this implies that such mutants can successfully invade the population.

Remark 3. The argument above suggests that in our setup evolutionary forces tend to eliminate heritable heterogeneity, and the population at any given time should be concentrated around a single type. One can explain heterogeneity in the levels of the observed biases in the population in a slightly richer version of our model in which each bias is partially heritable and its realized level depends also on random type-independent factors. For further discussion see the beginning of Section 6 below.

5 Main Results

This section studies how biases change over time under the hybrid-replicator dynamics.

5.1 Only the Rational Type is Asymptotically Stable

In a seminal paper, Waldman (1994) characterizes dynamically stable types under sexual inheritance. In this subsection we apply his characterization to the current setup and show that only the rational type is asymptotically stable.

We begin by defining asymptotic stability. Suppose that a population composed of only type-t agents is invaded by a small group of mutants of type t'. If the population (1) does not move far away from its pre-entry state, and (2) converges back to t, we say that type t is asymptotically stable against type t' . Type t is asymptotically stable if it is asymptotically stable against all types. Formally, let $g^{\tau}(\eta)$ be the induced distribution of type-τ generations after an initial distribution η (i.e, $g^2(\eta) = g(g(\eta))$, etc.).

Definition 5. Type $t \in T$ *is asymptotically stable* against type $t' \in T$ if for every $\lambda \in (0,1)$ there exists ϵ such that for every $\epsilon' \in (0,\epsilon)$:

1. (Lyapunov stability) $g^{\tau}(\epsilon' \cdot t' + (1-\epsilon') \cdot t)(t) > 1-\lambda$ for every $\tau \in \mathbb{N}$; and
2. $lim_{\tau\to\infty} g^{\tau}(\epsilon' \cdot t' + (1-\epsilon') \cdot t) = t$.

Type t is a single-bias Nash equilibrium if no other type that differs in a single trait is a best reply against t. Formally,

Definition 6. Type $t = (\chi,\psi) \in T$ *is a single-bias Nash equilibrium* in game $G_p = (T, u_p)$ if $u_p(t,t) \geq u_p(t',t)$ for all $t' = (\chi',\psi') \in T$ such that $\chi = \chi'$ *or* $\psi = \psi'$.

Observe that any type that is a Nash equilibrium (see Def. 3) is also a single-bias Nash equilibrium. Waldman (1994) deals with a simpler setup without strategic interactions, in which the type's payoff is independent of the population. In this setup: (1) a single-bias Nash equilibrium is interpreted as a *second-best adaptation*, which optimizes the type's payoff under the constraint that only one of the biases can be changed with respect to the incumbent type t, and (2) a Nash equilibrium is interpreted as *first-best adaptation*, which optimizes the type's payoff without constraints. Waldman (1994, Prop. 2) shows that being a single-bias Nash equilibrium is a necessary condition for asymptotic stability. Prop. 2 adapts Waldman's (1994) analysis to the current strategic setup and shows that for any sufficiently small $p > 0$, the rational type is (1) the unique single-bias Nash equilibrium, and (2) the unique asymptotically stable type.

Proposition 2. *Let $t \neq (0, I) \in T$. There exists $\bar{p} > 0$ such that for any $p \in (0, \bar{p})$: (1) type t is not a single-bias Nash equilibrium, and (2) type t is not asymptotically stable. By contrast, type $(0, I)$ is a strict Nash equilibrium and asymptotically stable for any $p \in [0, 1]$.*

Sketch of proof. (full proof appears in the Appendix)

1. We deal with two separate cases:

   (a) $t = (\chi, \psi) \notin \Gamma$: The two biases of type t do not cancel each other in the barter interaction, and thus, type t does not choose the optimal threshold against itself. This implies that there exists type $t' = (\chi, \psi')$ that differs from t only in the perception bias, and that the bias ψ' induces type t' to choose the optimal thresholds against a population of agents of type t. This implies that $u_0(t', t) > u_0(t, t)$, and the same inequality holds for u_p for a sufficiently small p. Thus type t is not a single-bias Nash equilibrium. As observed in Section 4.2, in a population where all agents have the same level of cursedness, the hybrid-replicator dynamics is payoff monotone, and thus mutants of type t' outperform incumbents of type t, which implies that type t is not asymptotically stable.

   (b) $t = (\chi, \psi) \in \Gamma \setminus \{(0, I)\}$: Consider a mutant type $t' = (\chi', \psi)$ with the same perception bias and a slightly lower level of cursedness $\chi' = \chi - \epsilon$. Agents of type t' choose slightly different thresholds than agents of type t do (specifically, the former tend to trade a bit less). By a standard "envelope theorem" argument, these slightly different thresholds have only a second-order effect on the agent's payoff in the barter trade (because type t's thresholds are optimal). By contrast, the slightly lower level of cursedness induces a positive first-order effect on the payoff in the additional activities. This implies that for any $p > 0$, if ϵ is sufficiently small, then $u_p(t', t) > u_p(t, t)$. This shows that type t is not a single-bias Nash equilibrium, and by the same argument as in Case (A), it is not asymptotically stable.

2. As noted above, the rational type always does strictly better than any other type, i.e., $u_p((I,0),\cdot) > u_p(t',\cdot)$. That is, the rational type is a strict Nash equilibrium for any $p > 0$. Consider any mutant type $t' = (\chi', \psi')$ with a small mass of ϵ that invades a population of $(I,0)$-agents. Almost all $(1 - O(\epsilon))$ of the agents with cursedness 0 also have perception bias I (and vice versa), and these rational agents achieve a strictly higher payoff than agents who have either cursedness χ' or bias ψ'. This implies that the frequency of cursedness 0 and perception I converge to one, which shows that $(I,0)$ is asymptotically stable.

□

Waldman (1994) shows that a single-bias Nash equilibrium is necessary for stability under sexual inheritance (and it is sufficient if it induces a payoff that is not too low). Waldman (1994) applies this result to a setup with a finite set of feasible types $\{(\chi_1,\psi_1),...,(\chi_N,\psi_N)\}$, in which many biased types are single-bias Nash equilibria. A biased type (ψ_i,χ_i) in which the two traits only partially compensate for each other is a single-bias Nash equilibrium because the set of feasible types is sufficiently sparse, such that no other ψ_j better compensates for χ_i (and, similarly, no other χ_j better compensates for ψ_i).

However, if the set of types is a continuum and the payoff function is concave, as in our setup, then only the rational type is a single-bias Nash equilibrium, and thus it is the unique asymptotically stable type.[22] As long as the payoff function is concave, this is also the case in an environment with a finite but sufficiently dense set of types. For example, consider a setup similar to ours except that the set of feasible types is discretized and Δ is the distance between any two neighboring types. If $p \ll \Delta$, then types sufficiently close to Γ will be single-bias Nash equilibria, that is will outperform any "single-bias" mutant who is different in one of the biases.. This is because the mutant's disadvantage in best-replying in the barter interactions will outweigh any advantage the mutant might have in non-barter interactions. However, if $\Delta \ll p$, then any type in (or near) Γ is unstable against nearby "single-bias" mutants with a slightly lower level of one bias. Arguably, in a biological setup in which the set of feasible values that can be encoded by the chromosomes is very large, the latter case may be plausible even for small values of p.

[22] As noted by Waldman (1994, page 488), if the set of feasible types is convex, and if the payoff function is concave (as in our setup), the set of Nash equilibria coincides with the set of single-bias Nash equilibria.

5.2 Global Fast Convergence to Γ and Slow Drift within Γ

The previous subsection applied an analysis à la Waldman (1994) and showed that only the rational type is asymptotically stable. In this subsection we present a novel analysis that goes beyond Waldman (1994) and shows (1) fast convergence to a small neighborhood around Γ from any initial type, and (2) the stability of the types in Γ against all types except for those that are close neighbors to Γ and have slightly smaller biases.

We begin by defining the L_1-norm to measure distance between types (the same results can be achieved by other standard norms, such as the L_2-norm). Formally, let the distance between two perception biases ψ, ψ' and the distance between two types $t = (\chi, \psi)$, $t' = (\chi', \psi')$ be defined as follows:

$$\|\psi - \psi'\| = \int_L^H |\psi(x) - \psi'(x)|\, dx, \qquad \|t - t'\| = |\chi - \chi'| + \|\psi - \psi'\| .$$

Let $\|t\| = \|t - (0, I)\|$ be the distance between t and the rational type. For a set of types $\tilde{T} \subseteq T$ and a type t' let $\left\|\tilde{T} - t'\right\| = inf_{t \in \tilde{T}} \|t - t'\|$. Finally, given type $t \in T$, let $t_\Gamma \in \Gamma$ be the type in Γ with the same level of cursedness as t.

For any $\delta > 0$, let $\Gamma_\delta = \{t \in T \mid \|\Gamma - t\| \leq \delta\}$ be the set of types that are δ-close to Γ. In what follows we show our first main result: for any type $t \notin \Gamma_\delta$ there is a mutant type t' strictly closer to Γ_δ that eliminates type t, i.e.; any initial population that includes both types converges to a population in which all agents have type t'. Moreover, there is a uniform bound $n = n(\delta)$, such that at most n sequential invasions can take the population from any initial state to Γ_δ. Formally:

Definition 7. Type $t' \in T$ *eliminates* $t \in T$ if $\forall\, \epsilon \in (0, 1)$: $lim_{\tau \to \infty} g^\tau(\epsilon \cdot t' + (1 - \epsilon) \cdot t) = t'$.

Proposition 3. *For each $\delta > 0$, there exists $\epsilon = \mathcal{O}(\delta) > 0$ and $\bar{p} > 0$, such that for each type $t \notin \Gamma_\delta$, there exists type t' such that: (1) type t' eliminates t for any $p \in (0, \bar{p})$ and any $r \in [0, 1]$, and (2) $\|t_\Gamma - t'\| < \|t_\Gamma - t\| - \epsilon$. This implies that there is a sequence of finite types $(t_0 = t, t_1, ..., t_n)$ of length $n \leq \frac{1}{\epsilon}$ that is independent of p, such that each type t_{i+1} eliminates type t_i, and type t_n is δ close to Γ, i.e., $t_n \in \Gamma_\delta$.*

The sketch of the proof is similar to part 1 (a) of Prop. 2. It relies on the elimination of any type $t \notin \Gamma_\delta$ by a type t' with the same level of cursedness as t, and a perception bias that induces type t' to choose the optimal thresholds against a population of agents of type t.[23] Observe that the upper bound on the number of sequential invasions that are required to

[23] Observe that Γ contains types with all possible levels of $\chi \in [0, 1]$ (see Equation 2). Thus, we can establish a convergence to Γ from any type t using mutants that differ only in their perception bias (which gets closer

converge to Γ_δ (namely, $\frac{1}{\epsilon}$) is independent of p (for any sufficiently small p) and is independent of the initial type.

Next, we show that any type $t \in \Gamma_\delta$ is stable against all types that are not in its ϵ-neighborhood (with $\epsilon = \mathcal{O}(p)$). This implies that the minimal length of a sequence of invasions, taking the population from type $t \in \Gamma_\delta$ to the rational type $(0, I)$, is at least $\frac{\|t\|}{\epsilon}$, and in particular, this minimal length converges to infinity as p converges to zero. The second part of the result shows that types in its ϵ-neighborhood with slightly smaller biases, eliminate type t. Formally:

Proposition 4. *Fix $r > 0$. For each $\bar{p} > 0$, there is $\epsilon = \mathcal{O}(\bar{p})$ such that:*

1. *there is $\bar{\delta} > 0$ such that for any $p \in (0, \bar{p})$ each type $t = (\chi, \psi) \in \Gamma_{\bar{\delta}}$ is asymptotically stable against every other type $t' = (\chi', \psi')$ satisfying $|\chi' - \chi|, \|\psi - \psi'\| > \epsilon$; this implies that the minimal length of a sequence $(t_0 = t, t_1, ..., t_n = (I, 0))$ such that each type t_{i+1} eliminates type t_i, must satisfy that $n \geq \frac{\|t\|}{\epsilon} = \|t\| \cdot \Omega\left(\frac{1}{\bar{p}}\right)$.*

2. *for any $p \in (0, \bar{p})$, there is $\delta > 0$ such that each type $t = (\chi, \psi) \in \Gamma_\delta \setminus \{(0, I)\}$ can be eliminated by a type t' satisfying $\|t - t'\| < \epsilon$ and $\|t\| - \epsilon < \|t'\| < \|t\|$.*

Sketch of proof.

1. Type $t \in \Gamma_\delta$ makes almost optimal decisions in the barter trade. For any mutant t' that is sufficiently far away from t, either t' itself, or its hybrid offspring (that inherits one bias from t' and one bias from t), chooses thresholds that are substantially different from the almost optimal thresholds of type t. This implies that the advantage of the incumbent type in the barter trade is substantial, and that it outweighs any possible loss in the additional activities for a sufficiently small p, which implies the asymptotic stability of type t'.

2. The sketch of the proof is similar to part 1 (b) of Prop. 2.

□

5.3 Summary of Results

Combining our results, we conclude that for a sufficiently small $p > 0$ the population dynamics satisfies the following properties:

to ψ_χ for a given χ). The space of perception biases is much richer, however, and one cannot (in general) converge to Γ by having mutants that differ only in their level of cursedness: types with $\psi(x) \notin \{\psi_\chi\}$ are not in Γ.

1. *Homogeneity of the population*: the informal argument in Section 4.3 suggests that any heterogeneous population converges to a homogeneous population as soon as a mutant type with biases close to the population's average type appears.

2. *Global convergence to* Γ (Prop. 3): any type far from Γ can be eliminated by types closer to Γ. Moreover, a finite sequence of invasions can take the population to be very close to Γ, and the bound on the length of this sequence is independent of p.

3. *"Punctuated" stability of types in* Γ (Prop. 4 – part 1): each type t close to Γ is asymptotically stable against all types except for those that are very close to t and have slightly smaller biases. This implies that each biased type close enough to Γ can survive as a unique incumbent type for a substantial period of time, and once it is eliminated, it will be replaced by a nearby slightly less biased type.

4. *Slow drift toward* $(0, I)$: the population slowly drifts toward $(0, I)$ in a small neighborhood around Γ (Prop. 4 – part 2 + "this implies" of part 1). Each step in this drift requires the appearance of a new mutant with slightly smaller biases, and the length of the sequence of invasions that eventually take the population all the way to $(0, I)$ is $\Omega\left(\frac{1}{p}\right)$.

6 Discussion

Random Traits and Empirical Predictions For simplicity, we have described agents with a particular type as having the same cursedness level and the same perception bias. However, the results remain qualitatively the same in an extended model in which the biases are only partially heritable, that is, if the cursedness level and the perception bias that an agent of type (χ, ψ) exhibits are $\chi + \epsilon_\chi$ and $\psi + \epsilon_\psi$, respectively, where ϵ_χ and ϵ_ψ are random factors unrelated to the agent's type and are independent of each other.

In Section 4.3 we concluded that evolutionary forces tend to eliminate heterogeneity, and the population at almost any given time should be concentrated around a single type in, or very close to, Γ. Thus, in this extended model, *any* observable heterogeneity within a given population should be the result of random non-heritable variations of the biases described in the previous paragraph. This leads to two empirical predictions, which allow, in principle, to test the extended model. First, the model predicts that, in a given population, there will be no correlation between the levels of the two biases. Second, the extended model predicts that on average these biases will approximately compensate for each other in barter-trade interactions, which were typical in the evolutionary history. It may also be argued that different sub-populations that started from different initial conditions have not yet had time to mix completely. Such sub-populations are expected to be at different states in Γ and thus have heterogeneous average levels of the two biases. If one can compare between such sub-populations, then the third prediction of this extended model is to see a strong positive correlation between the average cursedness level and the average endowment effect across sub-populations. Obviously, empirical research is needed in order to confirm or falsify these predictions, and such a process may be complicated.

Richer Environments The model described in Section 4.1 can capture (with slight modifications) an environment in which agents randomly engage in one of *many* barter-trade interactions that are relatively similar to each other (rather than playing one game with high probability and a completely different activity with low probability, as illustrated in Section 4). For concreteness, consider the following example. Assume that each period traders' private signals are drawn from a different distribution. The distribution is chosen i.i.d. each period according to a log-normal random variable $\nu_\sigma \sim \ln N\left(0, \sigma^2\right)$. Given ν_σ, the private value of each trader is distributed according to the CDF $F_{\nu_\sigma}(x) = \left(\frac{x-L}{H-L}\right)^{\nu_\sigma}$. Both agents publicly observe the realization of ν_σ before they decide whether or not to trade (the case in which ν_σ is unobservable is equivalent to playing a fixed barter trade). This environment can be embedded in our model. Define the function u as the payoff function in the barter trade when the distribution of private signals is $F_0(x) = \frac{x-L}{H-L}$, and define the function $v(\sigma)$ as the difference between the expected payoff (with respect to ν_σ) of the game with $F_{\nu_\sigma}(x)$, and u. The expected payoff in such an environment can be written as $u_\sigma = u + v(\sigma)$. Note that the standard deviation σ replaces p as the magnitude of the perturbation in the model. Observe that: (1) the function $u_\sigma(t, \eta)$ describes the expected payoff of type t in population η in this environment; (2) $v(\sigma) \sim O(\sigma)$ for a sufficiently small σ; and (3) function $v(\sigma)(t, \eta)$ is Lipschitz continuous and bias monotone. Due to these observations, all our stability results can be extended to this setup in a relatively straightforward way when σ is sufficiently small. This example can be generalized to any family of ex-ante distributions of private values, as long as the variance among these distributions is sufficiently small.

Different Ways to Capture the Biases Our qualitative results are robust to the exact way in which the two biases are captured in the model. One can capture cursedness by following a different approach than Eyster & Rabin (2005) (e.g., by adopting the "analogy based expectation equilibrium" approach, as in Jehiel, 2005; Jehiel & Koessler, 2008). Since we allow a large set of perception bias functions, we can find a set of types for which the biases fully compensate each other also for this alternative representation of cursedness. One can claim that not every perception bias function is plausible, because biased individuals behave in a "simple" manner (for example, one can claim that ψ's must be linear). However, even for a large class of restricted sets of perception bias functions, where the two biases cannot fully compensate for each other, the results of Section 4 continue to hold. This is due to the fact that under the hybrid-replicator dynamics biases survive for long times even if they only approximately compensate for each other.

Different Barter-Trade Mechanisms Our results remain qualitatively similar to other variants of barter with indivisible goods, such as having interdependent values instead of common values, or having a different surplus coefficient for each trader. One can also consider trade with divisible goods in which the relative price of exchange between the commodities is endogenously determined by haggling. Such trading mechanisms turn out to be analytically intractable (though numeric analysis suggests that similar results may hold in such a setup). The analysis can be rendered tractable by assuming that the price is determined by a noisy double auction: if the buyer's price is lower than the seller's price no trade occurs, whereas if it is higher then the probability of trade is increasing linearly in the difference between the two bids, and the price is the mean of the two bids. In this specific setup the results remain qualitatively similar.

7 Conclusion

This paper explores the possibility that several biases coevolve together if they approximately compensate for each other errors, and thus lead to behavior that is "close" to a fitness maximizing ("rational") behavior. By focusing on barter trade, an activity of significant importance in prehistoric times, we have specifically suggested that the Endowment Effect and Winner's Curse (Cursedness) may have coevolved. We have also shown that even when biases lead to fitness that is strictly lower than the fitness of rational types, biases can survive for a long time, as long as the evolutionary dynamic is non-monotone. Our main methodological innovation is to show that this is true even when there are no pairs of biases that are "second-best adaptations" (i.e., there exist no pairs of biases in which the the level of each bias is optimal when taking the level of the other bias as fixed). We hope our results can be used in the future to deepen the understanding of the evolution of behavioral biases.

A Technical Appendix

A.1 Proof of Lemma 1 (Existence of Equilibrium Configurations)

Denote by F the absolutely continuous cumulative distribution function of $\mathbf{x}_i$. The explicit formula for the expected value of the partner's good conditional on his agreement to trade , $\mu_\alpha(\eta, b)$, is

$$\mu_\alpha(\eta, b) = \frac{\sum_{t \in \text{supp}(\eta)} \eta(t) \cdot F(b(t)(\alpha)) \cdot \mu_{\le b(t)(\alpha)}}{\sum_{t \in \text{supp}(\eta)} \eta(t) \cdot F(b(t)(\alpha))}, \tag{5}$$

and if the denominator equals 0 (i.e., no agent ever agrees to trade given coefficient α), let $\mu_\alpha(\eta, \beta) = L$.

Let $\{t_i = (\chi_i, \psi_i)\}_{i \le n}$ be the finite set of types in population η (n is the arbitrary number of types). Substituting (5) into (1), configuration (η, b) is an equilibrium if and only if for each type $i \le n$, whenever $\alpha \cdot (\chi_i \cdot \mu + (1 - \chi_i) \mu_\alpha(\eta, b)) \le \psi_i(H)$,

$$\psi_i(b(t_i)(Foreachr, \epsilon > 0, thereis\bar{p} = O(\min(\epsilon, r))\alpha)) = \alpha \cdot (\chi_i \cdot \mu + (1 - \chi_i) \cdot \mu_\alpha(\eta, b)), \tag{6}$$

and $b(t_i)(\alpha) = H$ otherwise.

Fix any $\alpha \in \left[1, \frac{H}{L}\right]$. Let $x_i = b(t_i)(\alpha)$ and $\eta_i = \eta(t_i)$. Then (6) is reduced to the following set of n equations:

$$\forall i \le n \;\; \psi_i(x_i) = \alpha \left(\chi_i \cdot \mu + (1 - \chi_i) \cdot \frac{\sum_{j \le n} \eta_j \cdot F(x_j) \cdot \mu_{\le x_j}}{\sum_{j \le n} \eta_j \cdot F(x_j)} \right). \tag{7}$$

For each $i \le n$, let $g_{i,\alpha} : [L, \mu] \to [L, \alpha \cdot \mu]$ be a function that assigns the threshold of an agent of type t_i to any expected value of an object that is traded in population, v. Formally:

$$g_{i,\alpha}(v) = \begin{cases} \psi_i^{-1}(\alpha \cdot (\chi_i \cdot \mu + (1 - \chi_i) \cdot v)) & \alpha \cdot (\chi_i \cdot \mu + (1 - \chi_i) \cdot v) \le \psi_i(H) \\ H & \alpha \cdot (\chi_i \cdot \mu + (1 - \chi_i) \cdot v) > \psi_i(H). \end{cases} \tag{8}$$

Notice that in equilibrium $x_i = b(t_i)(\alpha) = g_{i,\alpha}(\mu_\alpha(\eta, b))$. Now let $h_\eta : [L, \alpha \cdot \mu]^n \to [L, \mu]$ be the function that assigns the expected value of an object that is traded in a population η to a profile of thresholds $(x_1, ..., x_n)$that are used in that population. Formally:

$$h_\eta(x_1, ..., x_n) = \frac{\sum_{j \le n} \eta_j \cdot F(x_j) \cdot \mu_{\le x_j}}{\sum_{j \le n} \eta_j \cdot F(x_j)}.$$

Notice that $\mu_\alpha(\eta, b) = h_\eta[b(t_1)(\alpha), ..., b(t_n)(\alpha)]$. Finally, let $f : [L, \mu] \to [L, \mu]$ be defined as follows:

$$f_\alpha(v) = h_\eta(g_{1,\alpha}(v), ..., g_{n,\alpha}(v)).$$

Observe that any solution to the equation $v = f_\alpha(v)$ induces thresholds that can be used as part of an equilibrium when the coefficient surplus is equal to α; that is, $v = \mu_\alpha(\eta, b)$ and $b(t_i)(\alpha) = g_{i,\alpha}(v)$. Similarly, any equilibrium configuration (η, b) induces a solution to the equation $v = f_\alpha(v)$ for all values of α. We now show that there is at least one solution for $v = f_\alpha(v)$. First observe that since we assume that F and ψ are continuous functions then g, h, and f are continuous functions. Second, notice that for each $\alpha \in \left[1, \frac{H}{L}\right]$ $f_\alpha(L) \geq L$ and $f_\alpha(\mu) \leq \mu$. Thus $f_\alpha(v) = v$ at some $v \in [L, \mu]$.

Next, we have to show that there exists a *continuous* function $v^* : \left[1, \frac{H}{L}\right] \to [L, H]$ such that for each $\alpha \in \left[1, \frac{H}{L}\right]$, we get[24] $f_\alpha(v^*(\alpha)) = v^*(\alpha)$. Let $v^*(\alpha)$ be the largest fixed point of f_α (i.e, for each α, we get $v^*(\alpha) = \arg\max_{v \in \left[1, \frac{H}{L}\right]} f_\alpha(v) = v$). The fact that ψ^{-1} and h are continuous and increasing functions as well as the linear dependency in α in Equation (8) imply (1) the continuity of $v^*(\alpha)$ with respect to α, and (2) the continuity of $v^*(\alpha)(\eta)$ with respect to the distribution of types η,.

Note that the above $v^*(\alpha)$ is the function that selects the equilibria with maximal trade. All of our results remain the same if one chooses a different continuous function $v^*(\alpha)$, such as the one that selects the smallest fixed point of f_α (and minimizes the probability of trade).

Finally, we have to show that in any equilibrium configuration (η, b) each strategy $b(t)$ is strictly increasing in α for each type $t \in supp(\eta)$. We prove it by the following steps:

1. Substituting $\alpha = \frac{H}{L}$ into (6) implies that the RHS is weakly larger than H, which implies that $b(t)\left(\frac{H}{L}\right) = H$ for any type $t \in supp(\eta)$ in any equilibrium configuration.

2. The function $\mu_\alpha(\eta, b)$ is a strictly increasing function of α in any equilibrium configuration. Otherwise, due to the continuity of the strategies $b(t_i)$ and the fact that $b(t_i)\left(\frac{H}{L}\right) = H$, there is $\alpha < \alpha'$, such that $\mu_\alpha(\eta, b) = \mu_{\alpha'}(\eta, b)$. Eq. (6) and the monotonicity of each ψ_i imply that $(b(t_i)(\alpha)) < (b(t_i)(\alpha'))$. Together with the equality $\mu_\alpha(\eta, b) = h_\eta[b(t_1)(\alpha), ..., b(t_n)(\alpha)]$, the former inequality implies that $\mu_\alpha(\eta, b) < \mu_{\alpha'}(\eta, b)$, a contradiction.

3. Because (i) the functions $\mu_\alpha(\eta, b)$ and $g_{i,\alpha}(v)$ are strictly increasing in α, and (ii) the functions $\psi_i(x)$ are strictly increasing, then $b(t_i)$ is strictly increasing in α.

[24] We have to show the continuity of $v^*(\alpha)$ in order to have all types using a continuous threshold strategy as a function of a surplus coefficient (as assumed in the definition of a strategy in Section 2.3).

A.2 Formal Definition of the Payoff Function of G_0

In this section we explicitly state the payoff of each type in the population game G_0. Let $u\left(s_1, s_2 | \alpha, x_1, x_2\right)$ be trader 1's payoff when his signal is x_1 and he plays s_1, while his partner, trader 2, has a signal x_2 and plays s_2; and when the public signal is α:

$$u_0\left(s_1, s_2 | \alpha, x_1, x_2\right) = \begin{cases} \alpha \cdot x_2 & x_1 \leq s_1\left(\alpha\right) \text{ and } x_2 \leq s_2\left(\alpha\right) \\ x_1 & \text{otherwise.} \end{cases}.$$

Now, let $u\left(s_1, s_2\right)$ be the expected payoff of an agent with strategy s_1 who faces a partner with strategy s_2, where the expectation is taken w.r.t. the values of of the signals $\mathbf{x_1}, \mathbf{x_2}$, and α:

$$u\left(s_1, s_2\right) = \int_{\alpha=1}^{H/\mu} \int_{x_1=L}^{H} \int_{x_2=L}^{H} u_0\left(s_1, s_2 | \alpha, x_1, x_2\right) \mathrm{d}F_{\mathbf{x_2}} \, \mathrm{d}F_{\mathbf{x_1}} \, \mathrm{d}F_{\alpha},$$

where F_j is the CDF of random variable $i \in \{\alpha, \mathbf{x_1}, \mathbf{x_2}\}$. Finally, given type $t \in T$ and an equilibrium configuration $\left(\eta, b^*\left(\eta\right)\right)$, define $u\left(t, \eta\right)$ as the expected payoff of a type-$t \in T$ agent who faces an opponent randomly selected from population η:

$$u\left(t, \eta\right) = \sum_{t' \in supp(\eta)} \eta\left(t'\right) \cdot u\left[s_t^*\left(\alpha\right)\left(\eta, b^*\left(\eta\right)\right), b^*\left(\eta\right)\left(t'\right)\right].$$

Note that the definition is well defined also for types outside the support of η (i.e., for types $t \in T \backslash supp\left(\Gamma\right)$).

A.3 Proof of Proposition 1 (Characterization of Equilibria in G_0)

The proof includes the following parts:

1. Definitions and notations:

   (a) *Probability of trade*: let $q\left(\alpha | \eta\right)$ be the probability that a random partner from population η agrees to trade given α. Note, that $q\left(\alpha | \eta\right) > 0$ iff $\exists t \in$ *supp* $\left(\eta\right)$ with $b_\eta^*\left(t\right)\left(\alpha\right) > L$.

   (b) *Expected payoff of a threshold strategy*: let $u\left(x, \alpha | \eta\right)$ be the expected payoff of a player who uses threshold x, and faces a distribution of types η (which plays according to b_η^*), conditional on the surplus coefficient being α.

   (c) *Incumbents*: the types in the support of a given distribution η.

(d) *A type's threshold strategy*: for each type $t = (\chi, \psi)$ let strategy $s_t^*(\alpha|\eta)$ be the threshold strategy of a type t who faces population η; that is, for each $\alpha \in \left[1, \frac{H}{L}\right]$, $s_t^*(\alpha|\eta)$ is the unique solution to the equation:

$$\psi\left(s_t^*(\alpha|\eta)\right) = \min\left\{\alpha \cdot \left(\chi \cdot \mu + (1-\chi) \cdot \mu_\alpha\left(\eta, b_\eta^*\right)\right), H\right\}.$$

(e) *Mean threshold*: For each $\alpha \in \left[1, \frac{H}{L}\right]$ and distribution of types $\eta \in \Delta(T)$, let $\bar{x}(\alpha|\eta)$ be the unique solution to the equation: $\mu_{\leq \bar{x}(\alpha|\eta)} = \mu_\alpha(\eta)$; that is, for a given α, the mean value of a traded good in a homogeneous population where all agents use threshold $\bar{x}(\alpha|\eta)$ is equal to the mean value of a traded good in population η (where all agents play equilibrium strategies).

2. The expected payoff of a threshold is given by the following formula:

$$u(x, \alpha|\eta) = u(\alpha \cdot \mu_\alpha, \alpha|\eta) - q(\alpha|\eta) \cdot |F(x) - F(\alpha \cdot \mu_\alpha)| \cdot E\left(|y - \alpha \cdot \mu_\alpha| \mid y \in [\alpha \cdot \mu_\alpha, x]\right), \tag{9}$$

where, with a slight abuse of notation, $[\alpha \cdot \mu_\alpha, x] = [x, \alpha \cdot \mu_\alpha]$ if $x < \alpha \cdot \mu_\alpha$. The *optimal threshold*, which induces the maximal payoff, is $x = \alpha \cdot \mu_\alpha$ because it results in trade iff the trader's good is worth less than the expected value of a trading partner. Using a different threshold x yields a wrong decision with probability $q(\alpha|\eta) \cdot |F(x) - F(\alpha \cdot \mu_\alpha)|$. Conditional on making a wrong decision and the partner's agreement to trade, the expected loss from trade is equal to $E(|y - \alpha \cdot \mu_\alpha| \mid y \in [\alpha \cdot \mu_\alpha, x])$ (the expected difference between the value of the trader's good and the conditional expected value of his partner). Observe that $u(x, \alpha|\eta)$ is concave w.r.t. x, and strictly concave if $q(\alpha|\eta) > 0$.

3. The previous step immediately implies that type $(0, I)$ (who always chooses the optimal threshold of $\alpha \cdot \mu_\alpha$) weakly outperforms any other type (i.e., $u((0, I), \eta) \geq u(t, \eta)$ for each $t \in T$ and $\eta \in \Delta(T)$), and strictly outperforms any other type if there is any α such that $q(\alpha|\eta) > 0$ and $x \neq \alpha \cdot \mu_\alpha$.

4. The "if" side. Let $\eta \in \Delta(T)$ be a Nash equilibrium of G_0. We prove that *supp* $(\eta) \in \Gamma$.

(a) In any Nash equilibrium all types use the optimal thresholds for all α's; that is, $\forall t \in supp(\eta)$ and $\alpha \in \left[1, \frac{H}{L}\right]$ $b_\eta^*(t)(\alpha) = \alpha \cdot \mu_\alpha(\eta)$.

Assume to the contrary that there is at least one value of α for which one of the incumbent types uses a non-optimal threshold. This set of α's is defined as

$$A_\alpha = \left\{\alpha \mid \exists t_0 \in supp(\eta) \text{ s.t. } b_\eta^*(t_0)(\alpha) \neq \alpha \cdot \mu_\alpha(\eta)\right\}.$$

By part 3 and the continuity of the threshold strategies, an incumbent type in a Nash equilibrium distribution can use a non-optimal threshold in α only if $q\left(\alpha|\eta\right) = 0$.

Let $\bar{\alpha}$ be the supremum of A_{α}. Consider first the case where $\bar{\alpha} = \frac{H}{L}$. In such a case, the assumptions that each ψ is strictly increasing and $\psi\left(H\right) \leq H$ imply that $\psi\left(x\right) < H$ for each $x < H$, and therefore $q\left(\alpha|\eta\right) > 0$ for α's sufficiently close to $\frac{H}{L}$, and $u\left(\left(0, I\right), \eta\right) > u\left(t, \eta\right)$ for some type $t \in supp\left(\eta\right)$.

Now consider the case where $\bar{\alpha} < \frac{H}{L}$. All types play the rational threshold for $\alpha > \bar{\alpha}$, and by continuity of $b_{\eta}^{*}\left(t\right)\left(\alpha\right)$ all agents play the rational threshold at $\bar{\alpha}$. Observe that the rational thresholds are always strictly larger than L, and this implies that $q\left(\alpha|\eta\right) > 0$ for each $\alpha \geq \bar{\alpha}$. By continuity of $b_{\eta}^{*}\left(t\right)\left(\alpha\right)$, there exists an interval of α's such that $\alpha < \bar{\alpha}$, $q\left(\alpha|\eta\right) > 0$, and $\alpha \in A_{\alpha}$, and this implies that η cannot be a Nash equilibrium.

(b) If there exists $t_0 = \left(\chi_0, \psi_0\right) \in T \backslash \Gamma$ in $supp\left(\eta\right)$, then there exists an α such that type t_0 does not use the optimal threshold; i.e., $\exists \alpha_0 \in \left[1, \frac{H}{L}\right] \;\; b_{\eta}^{*}\left(t_0\right)\left(\alpha\right) \neq \alpha \cdot \mu_{\alpha}\left(\eta\right)$.

Assume to the contrary that for each $\alpha \in \left[1, \frac{H}{L}\right]$, all incumbents $t \in supp\left(\eta\right)$ use the "optimal" threshold $b_{\eta}^{*}\left(t\right)\left(\alpha\right) = x^{*}(\alpha) \equiv \min\left\{\alpha \cdot \mu_{\alpha}, H\right\}$. In the next argument, we focus on the interval of α in which $x^{*}(\alpha)$ is determined by the indifference condition (1). It is easy to see that such values always exist for types who trade optimally. For these levels of α, we obtain

$$\psi_0\left(\alpha \cdot \mu_{\alpha}\right) = \alpha \cdot \left(\chi_0 \cdot \mu + \left(1 - \chi_0\right) \cdot \mu_{\alpha}\right).$$

The fact that all players use the optimal thresholds implies that $\mu_{\alpha} = \mu_{\leq x^{*}(\alpha)}$, and therefore $x^{*}(1) = L$ and $x^{*}\left(\frac{H}{\mu}\right) = H$. By continuity, $x^{*}\left(\alpha\right)$ obtains all values in $[L, H]$, and note that $x^{*}\left(\alpha\right) = H$ for $\alpha \in \left[\frac{H}{\mu}, \frac{H}{L}\right]$. Given $x^{*}(\alpha) \equiv \alpha \cdot \mu_{\alpha}$, we can rewrite the indifference condition above as follows: $\forall x^{*} \in [L, H]$,

$$\psi_0\left(x^{*}\right) \;\;=\;\; \alpha \cdot \left(\chi_0 \cdot \mu + (1 - \chi) \cdot \frac{x^{*}}{\alpha}\right) = \chi_0 \cdot \frac{\mu}{\mu_{\leq x^{*}}} \cdot x^{*} + (1 - \chi) \cdot x^{*} = \psi_{\chi_0}^{*}(x^{*}),$$

which implies that $\psi_0(x) = \psi_{\chi_0}^{*}(x)$ – contradicting the assumption that $t_0 \in T \backslash \Gamma$.

5. The "only if" side and the "moreover" statement. Let η_0 be a distribution with $supp\left(\eta_0\right) \in \Gamma$. We prove that η_0 is a Nash equilibrium, that all incumbents use the same threshold strategy, and that each type $t' \notin \Gamma$ is strictly outperformed (i.e, $u\left(t', \eta\right) < u\left(t_0, \eta\right)$ for each $t_0 \in supp\left(\eta_0\right)$).

(a) In Section 3.1 we showed that if all other agents use the "rational" threshold that is defined by $x^*(\alpha) = \min\left\{\alpha \cdot \mu_{<x^*(\alpha)}, H\right\}$, then a type $t \in \Gamma$ also uses $x^*(\alpha)$. Thus, $b^* = x^*(\alpha)$ is an equilibrium behavior induced by distribution η_0. In what follows we show that there are no other equilibrium behaviors induced by distribution η_0.

(b) Assume to the contrary another equilibrium behavior that induces $\bar{x}(\alpha|\eta_0) \neq x^*(\alpha)$:

i. If $\bar{x}(\alpha|\eta) > \alpha \cdot \mu_\alpha(\eta_0)$ whenever $\alpha \cdot \mu_\alpha(\eta_0) < H$ then $s_t^*(\alpha|\eta_0) \leq \bar{x}(\alpha|\eta_0)$ for each $t = \left(\chi, \psi_\chi^*\right) \in \Gamma$. To see why, assume that α is such that $\alpha \cdot \mu_\alpha(\eta_0) < H$. Recall that ψ_χ^* is strictly increasing, and that $s_t^*(\alpha|\eta_0)$ is the unique solution to the following equation:

$$\begin{aligned}\psi_\chi^*\left(s_t^*(\alpha|\eta_0)\right) &= \alpha \cdot \left(\chi \cdot \mu + (1-\chi) \cdot \mu_\alpha(\eta_0)\right) \\ &< \chi \cdot \frac{\mu}{\mu_{\leq \bar{x}(\alpha|\eta_0)}} \cdot \bar{x}(\alpha|\eta_0) + (1-\chi) \cdot \bar{x}(\alpha|\eta_0) = \psi_\chi^*\left(\bar{x}(\alpha|\eta_0)\right),\end{aligned}$$

where the strict inequality is implied by $\bar{x}(\alpha|\eta_0) > \alpha \cdot \mu_\alpha(\eta_0)$ and $\mu_\alpha(\eta_0) = \mu_{\leq \bar{x}(\alpha|\eta_0)}$. Since $\bar{x}(\alpha|\eta)$ is the average threshold, we get a contradiction.

ii. By an analogous argument, if $\bar{x}(\alpha|\eta_0) < \alpha \cdot \mu_\alpha$ then $s_t^*(\alpha|\eta_0) > \bar{x}(\alpha|\eta_0)$ for each $t = \left(\chi, \psi_\chi^*\right) \in \Gamma$, and again we get a contradiction.

iii. Therefore, it must be that $\bar{x}(\alpha|\eta_0) = \alpha \cdot \mu_\alpha$ whenever $\alpha \cdot \mu_\alpha(\eta_0) < H$. In such a case, by an analogous argument, $s_t^*(\alpha|\eta_0) = \bar{x}(\alpha|\eta_0) = x^*(\alpha)$ for each $t = \left(\chi, \psi_\chi^*\right) \in \Gamma$.

(c) The previous parts imply that if *supp*$(\eta_0) \in \Gamma$, then all incumbent types use threshold strategy $x^*(\alpha)$ and are internally equivalent. Moreover, this threshold strategy is the optimal one, and this implies that η_0 is a Nash equilibrium.

(d) Finally, we can use a similar argument to part (4b) to show that for any α there is a positive probability of trade, and therefore any type $t' \notin \Gamma$ that uses a non-optimal threshold for some α's has a strictly lower payoff than the incumbents.

A.4 Asymptotic Stability of Γ in the Game G_0

Prop. 1 in Section 3.2 shows that the set Γ is an internally equivalent and externally strict set. In this appendix we show why these properties imply that the set of distributions over Γ is asymptotically stable in the replicator dynamics.

Suppose a dynamic game in which agents are randomly matched each period (two independent draws from η) and play the trade game. Agents' payoffs determine their fitness and therefore their frequency in the population in the next stage; that is, the type distribution η evolves through a payoff-monotone selection dynamics, such as the replicator dynamics (Taylor & Jonker, 1978). Given Proposition 1, we can rely on existing results to relate the static equilibria with dynamically stable distributions in such a dynamics. We sketch below the argument why the set Γ is dynamically stable, and the types outside Γ are unstable.

A distribution of types η is *Lyapunov stable* if after any sufficiently small invasion by a mutant distribution of types, the population composition remains close to η in all future generations.[25]A set of Lyapunov-stable distributions is *asymptotically stable* if, after any small enough invasion by a mutant distribution to any of the distributions in the set, the population reverts back to the set in the long run. An asymptotically stable set is *minimal* if no strict subset is asymptotically stable. For brevity we omit the formal definitions and the formal arguments, which are quite standard.

Recall that we restrict attention to distributions of types with finite support (both for incumbents and for mutants), and thus we can apply the result by Nachbar (1990) that any Lyapunov-stable distribution is a Nash equilibrium. Thomas (1985) defines a notion of an *evolutionarily stable set* and shows that it implies asymptotic stability in the replicator dynamics in a finite strategy space (Cressman, 1997, extends this result to a large set of payoff-monotone dynamics). Norman (2008, Theorem 1) further extends Thomas's result to infinite strategy spaces. Specifically, he shows that if the set of all distributions over a Borel set is evolutionary stable with a uniform invasion barrier, then it is asymptotically stable.

Let $\Delta_\Gamma \subset \Delta(T)$ be the set of distributions over Γ. The fact that Γ is internally equivalent and externally strict implies that Δ_Γ is evolutionarily stable. It is also relatively straightforward to see that the fully cursed type $(1, \psi_1^*) \in \Gamma$ has a uniform invasion barrier (i.e., there is $\bar{\epsilon} > 0$ such that the incumbent type $(1, \psi_1^*)$ strictly outperforms any mutant type outside Γ with mass $0 < \epsilon < \bar{\epsilon}$,) and that this invasion barrier also holds for any other distribution in Δ_Γ. Thus, one can adapt the result of Norman (2008) to the current setup and conclude that the set Δ_Γ is asymptotically stable.[26] This implies the following corollary of Proposition 1 that characterizes stable distributions in the replicator dynamics.

[25] We consider only perturbations in which an incumbent population is invaded by a small group of mutants (that is, we consider only the variational norm when assessing relevant nearby perturbations; see, e.g., Bomze, 1990). We do not consider "continuous" perturbations in which a large group of incumbents slightly change their types, as in Eshel & Motro (1981). This is because (1) a coordinated change in the type of many incumbents seems less plausible in our setup, and (2) the existing results on "continuous" stability (e.g., Oechssler & Riedel, 2002) hold only when the set of strategies is a subset of $\mathbb{R}$. By contrast, an analysis with the set of strategies in our population game is intractable.

[26] Norman (2008) formally deals with strategy spaces that are subsets of $\mathbb{R}^n$, but it seems that all the arguments in his proof can be extended to the current setup.

Corollary 1. *In game G_0 with an underlying replicator dynamics, (1) a distribution η is Lyapunov stable iff $\eta \in \Delta_\Gamma$, and (2) the set Δ_Γ is a minimal asymptotically stable set.*

A.5 Formal Definition of Hybrid-Replicator Dynamics

In this section we formalize the definition of the transition function $g : \Delta(T) \to \Delta(T)$ that is described informally in Section 4.2. The relative fitness $f_\eta(t)$ of type t in population η is

$$f_\eta(t) = \frac{\phi + u_p(t,\eta)}{\phi + E\left(u_p(t,\eta)\right)} = \frac{\phi + u_p(t,\eta)}{\phi + \sum_{t' \in supp(\eta)} \eta(t') \cdot u_p(t',\eta)}, \tag{10}$$

where $\phi \geq 0$ is the background expected number of offspring for an individual (unrelated to his payoff in the population game). Let $\mathcal{X}_\eta$ and Ψ_η be the cursedness levels and the perception biases in population η:

$$\mathcal{X}_\eta = \left\{\chi \in [0,1] \mid \exists \psi \in \Psi \text{ s.t. } (\chi,\psi) \in supp(\eta)\right\},$$

$$\Psi_\eta = \left\{\psi \in \Psi \mid \exists \chi \in [0,1] \text{ s.t. } (\chi,\psi) \in supp(\eta)\right\}.$$

For each $\chi \in \mathcal{X}_\eta$ ($\psi \in \Psi_\eta$) define $\eta(\chi)$ ($\eta(\psi)$) as the total frequency of types with χ (ψ):

$$\eta(\chi) = \sum_{\psi \in \ominus_\eta} \eta((\chi,\psi)) \quad \left(\eta(\psi) = \sum_{\chi \in \mathcal{X}_\eta} \eta((\chi,\psi))\right),$$

and define $f_\eta(\chi)$ ($f_\eta(\psi)$) as the mean relative fitness of types with cursedness χ (bias ψ):

$$f_\eta(\chi) = E\left(f_\eta((\chi,\psi)) \,|\psi \in \Psi_\eta\right) = \sum_{\psi \in \ominus_\eta} \frac{\eta((\chi,\psi))}{\eta(\chi)} \cdot f_\eta((\chi,\psi))$$

$$\left(f_\eta(\psi) = E\left(f_\eta((\chi,\psi)) \,|\chi \in \mathcal{X}_\eta\right) = \sum_{\chi \in \mathcal{X}_\eta} \frac{\eta((\chi,\psi))}{\eta(\psi)} \cdot f_\eta((\chi,\psi))\right).$$

Lastly, (for every $(\chi,\psi) \in \mathcal{X}_\eta \times \Psi_\eta$, the transition function is

$$g(\eta)((\chi,\psi)) = (1-r) \cdot \eta((\chi,\psi)) \cdot f_\eta((\chi,\psi)) + r \cdot \eta(\chi) \cdot f_\eta(\chi) \cdot \eta(\psi) \cdot f_\eta(\psi).$$

A.6 Lemma on Stability of Types in Hybrid-Replicator Dynamics

In what follows, we prove a lemma that characterizes stability in hybrid-replicator dynamics, and that is used in the proof of Propositions 2 and 4. The lemma shows that an incumbent type is asymptotically stable against a mutant type if: (1) the mutant's payoff is not substantially higher than the incumbent's payoff (specifically, the mutant's fitness should be less than $\frac{1}{1-r}$ times the incumbent's fitness); and (2) the hybrid types (who have one trait of type t and one trait of type t') yield lower payoffs than the incumbent. The lemma is an adaptation of Prop. 2 of Waldman (1994) to the current strategic setup in which the payoff of a type depends also on the population. Parts (2)–(3) of the lemma show that being a single-bias Nash equilibrium is a necessary condition for a type to be asymptotically stable. Part (4) of the lemma implies that being a strict Nash equilibrium is a sufficient condition for asymptotic stability.

Lemma 2 (Characterization of Stability in Hybrid-Replicator Dynamics)**.** *Let $t_1 = (\chi_1, \psi_1)$ and $t_2 = (\chi_2, \psi_2)$ denote some arbitrary types. Assume a population game G_p with a hybrid-replicator dynamics with parameters ϕ and r, and let $u_p(t, \eta)$ and $f_\eta(t)$ be the expected payoff and the relative fitness of type t against population η as defined in* (3) *and* (10)*.*

1. *If $(1-r) \cdot f_{t_1}(t_2) > 1$, then type t_1 is not asymptotically stable against t_2.*

2. *If $u_p((\chi_2, \psi_1), t_1) > u_p(t_1, t_1)$, then type t_1 is not asymptotically stable against (χ_2, ψ_1).*

3. *If $u_p((\chi_1, \psi_2), t_1) > u_p(t_1, t_1)$, then type t_1 is not asymptotically stable against (χ_1, ψ_2).*

4. *If **(a)** $(1-r) \cdot f_{t_1}(t_2) < f_{t_1}(t_1)$, **(b)** $u_p((\chi_2, \psi_1), t_1) < u_p(t_1, t_1)$, and **(c)** $u_p((\chi_1, \psi_2), t_1) < u_p(t_1, t_1)$, then type t_1 is asymptotically stable against t_2.*

Proof. Let $\epsilon > 0$ be sufficiently small, and let the initial distribution be: $\eta_0 = (1-\epsilon) \cdot t_1 + \epsilon \cdot t_2$.

Part (1). Assume that $(1-r) \cdot f_{t_1}(t_2) = c > 1$. By neglecting components that are $O(\epsilon^2)$, Eq. (4) implies that $g^\tau(\eta_0)(t_2) \approx \epsilon \cdot ((1-r) \cdot f_{t_1}(t_2))^\tau = \epsilon \cdot c^\tau$ and this implies instability of t_1 against t_2.

Parts (2)–(3) are immediately implied by well-known results for the replicator dynamics and the observation that when the population includes a single cursedness level (or a single perception bias), then the hybrid-replicator dynamics coincides with a replicator dynamics.

Part (4). Assume inequalities (a)–(c). Inequalities (b) and (c) imply $f_{t_1}((\chi_2, \psi_1)) < 1$ and $f_{t_1}((\chi_1, \psi_2)) < 1$. Define a constant c such that

$$\max\{(1-r) \cdot f_{t_1}(t_2), f_{t_1}((\chi_2, \psi_1)), f_{t_1}((\chi_1, \psi_2))\} < c < 1.$$

Let $\eta_\tau = g^\tau(\eta_0)$ be the distribution after τ generations. By neglecting components that are $O(\epsilon^2)$, Eq. (4) implies that for every τ in which $\eta_\tau((\chi_1, \psi_2))$ and $\eta_\tau((\chi_2, \psi_1))$ are $O(\epsilon)$: $\eta_\tau(t_2) \approx \epsilon \cdot c^\tau$, which converges to 0 at an exponential rate. Assume to the contrary, that

$\eta_\tau((\chi_1,\psi_2))$ does not converge to zero at an exponential rate. Then, for a sufficiently large τ, $\eta_\tau(t_2) \ll \eta_\tau((\chi_1,\psi_2))$. The average relative fitness of all types with $\psi = \psi_2$, $f_{\eta_\tau}(\psi_2)$, is a mixed average of $f_{\eta_\tau}((\chi_1,\psi_2))$ and $f_{\eta_\tau}(t_2)$, and as the weight of the latter converges quickly to zero, the average converges to the former, that is, $f_{\eta_\tau}(\psi_2) \approx f_{\eta_\tau}((\chi_1,\psi_2))$. Let $\epsilon' = \eta_\tau((\chi_1,\psi_2))$. We obtain that

$$\eta_{\tau+1}((\chi_1,\psi_2)) \approx (1-r)\cdot\epsilon'\cdot f_{\eta_\tau}((\chi_1,\psi_2)) + r\cdot\epsilon'\cdot f_{\eta_\tau}(\psi_2) \approx \epsilon'\cdot f_{\eta_\tau}((\chi_1,\psi_2)) < \epsilon'\cdot c,$$

which implies that $\eta_\tau((\chi_1,\psi_2))$ converges to zero at an exponential rate. An analogous argument works for (χ_2,ψ_1). □

A.7 Proof of Prop. 2 (Only $(0,I)$ is Asymptotically Stable)

1. Let $t = (\chi,\psi) \neq (0,I)$. Parts (2–3) of Lemma 2 imply that a single-bias Nash equilibrium is a necessary condition for asymptotic stability. In what follows, we prove that there is $\bar{p} > 0$ such that type t is not a single-bias Nash equilibrium for any $p \in (0,\bar{p})$. We deal with two separate cases: (a) $t \notin \Gamma$, and (b) $t \in \Gamma$.

   (a) Assume that $t \notin \Gamma$. This assumption implies that type t does not choose the optimal threshold against itself in the barter trade (by the same argument as in the proof of Prop. 1). Observe that the optimal (i.e., payoff-maximizing) strategy against a population of agents of type t is $s_{BR(t)}(\alpha) = \alpha\cdot\mu_{\leq b(t)(\alpha)}$. Define $\psi_{BR(t|\chi)}$ as the perception that chooses the optimal thresholds given cursedness χ and a population of t-agents, i.e., $\psi_{BR(t|\chi)}$ satisfies for each $\alpha \in \left[1,\frac{H}{L}\right]$:

   $$\psi_{BR(t|\chi)}\left(\alpha\cdot\mu_{\leq b_t^*(t)(\alpha)}\right) = \alpha\cdot\left(\chi\cdot\mu + (1-\chi)\cdot\mu_{\leq b_t^*(t)(\alpha)}\right).$$

   By Lemma 1, the conditional expectation $\mu_{\leq b(t)(\alpha)}$ is strictly increasing and continuous in α, which implies there exists a strictly increasing and continuous $\psi_{BR(t|\chi)}$ that satisfies these equations, and it is uniquely defined for each $x \geq \mu_{\leq b(t)(1)}$. If $\mu_{\leq b(t)(1)} > L$, then we still have to define $\psi_{BR(t|\chi)}(x)$ for lower values of x. Any strictly increasing continuous function will do, but for concreteness we choose the following linear definition: $\psi_{BR(t|\chi)}(L) = L$, and for each $x \in \left(L,\mu_{\leq b(t)(1)}\right)$:

   $$\psi_{BR(t|\chi)}(x) = L + \frac{x-L}{\mu_{\leq b_t^*(t)(1)} - L}\cdot\left(\chi\cdot\mu + (1-\chi)\cdot\mu_{\leq b_t^*(t)(1)}\right).$$

   The fact that $BR(t) = \left(\chi,\psi_{BR(t|\chi)}\right)$ chooses the optimal thresholds against t

implies that $u_0(BR(t), t) > u_0(t,t)$. This implies that there is $\bar{p} > 0$ such that for each $p \leq \bar{p}$ $u_p(BR(t), t) > u_p(t,t)$, which shows that type t is not a single-bias Nash equilibrium. Part (3) of Lemma 2 implies that type t is not asymptotically stable.

(b) Assume that $t \in \Gamma$. This assumption implies that an agent of type t chooses the optimal threshold against a population of t-players (by the arguments in the proof of Prop. 1). Let $\chi' = \chi - \epsilon$ for $0 < \epsilon \ll 1$. Observe that type $t' = (\chi', \psi)$ chooses almost the same thresholds as type t, i.e., for each $\alpha \in \left[1, \frac{H}{L}\right]$:

$$b_t^*(t')(\alpha) = b_t^*(t)(\alpha) + O(\epsilon) = s_{opt(t)}(\alpha) + O(\epsilon).$$

Observe that for each α the decision of an agent of type t' (against a population of t-agents) is optimal in most cases, except for those in which the agent's value is in the small interval of size $O(\epsilon)$ between $b_t^*(t')(\alpha)$ and $b_t^*(t)(\alpha)$, and in this case the expected loss is bounded by $O(\epsilon)$. This implies that the payoff of type t' is only $O(\epsilon^2)$ away from the payoff of type t in the barter trade, i.e., $u_0(t', t) \geq u_0(t,t) - O(\epsilon^2)$. On the other hand, type t' earns a higher payoff of $\frac{\partial v}{\partial \chi}(\chi) \cdot O(\epsilon) + O(\epsilon^2)$ in the additional activities (with $\frac{\partial v}{\partial \chi}(\chi) > 0$). This implies that for each $p > 0$, there is a sufficiently small $0 < \epsilon \ll p$ such that $u_p(t', t) > u_p(t,t)$, which shows that type t is not a single-bias Nash equilibrium.

2. Let $t' = (\chi', \psi')$. The fact that type $(0, I)$ makes optimal decisions in the barter trade and achieves the best payoff in v implies that for each $p > 0$ and each distribution of types in the population η, $u_p((0, I), \eta) > u_p(t', \eta)$. In particular, it implies that $(I, 0)$ is a strict Nash equilibrium. Part (4) of Lemma 2 implies that $(0, I)$ is asymptotically stable.

A.8 Proof of Proposition 3 (Global Fast Convergence to Γ)

Let $t = (\chi, \psi) \notin \Gamma$. Recall the definition of $BR(t)$ in the proof of Prop. 2 above. Let $t_\Gamma = (\chi, \psi_\chi) \in \Gamma$ be the element in Γ with the same level of cursedness as t. The proof includes the following steps.

1. We begin by showing that the perception bias of $BR(t)$ is strictly between t and t_Γ. Specifically, we show that for each $\alpha \in \left[1, \frac{H}{L}\right]$, the threshold chosen by type $BR(t)$ is

between the thresholds chosen by types t and t_Γ, i.e.,

$$\begin{aligned}
b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)=b_t^*\left(t\right)\left(\alpha\right) &\quad\Rightarrow\quad b_t^*\left(t_\Gamma\right)\left(\alpha\right)=b_t^*\left(t\right)\left(\alpha\right),\\
b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)<b_t^*\left(t\right)\left(\alpha\right) &\quad\Rightarrow\quad b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)\in\left(b_t^*\left(t_\Gamma\right)\left(\alpha\right),b_t^*\left(t\right)\left(\alpha\right)\right),\ \text{ and}\\
b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)>b_t^*\left(t\right)\left(\alpha\right) &\quad\Rightarrow\quad b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)\in\left(b_t^*\left(t\right)\left(\alpha\right),b_t^*\left(t_\Gamma\right)\left(\alpha\right)\right).
\end{aligned}$$

If $b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)=b_t^*\left(t\right)\left(\alpha\right)$, then it implies that type t chooses the optimal threshold against itself given α. Recall (see Section 3.1) that ψ_χ is defined to choose the optimal threshold against a population of agents who choose the optimal threshold (henceforth, an as-if rational population). This implies that $b_t^*\left(t_\Gamma\right)\left(\alpha\right)=b_t^*\left(t\right)\left(\alpha\right)$. Next, assume that $b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)<b_t^*\left(t\right)\left(\alpha\right)$ ($b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)>b_t^*\left(t\right)\left(\alpha\right)$). This implies that agents of type t choose a threshold that is too high (low) against a population of t-agents, and the expected value of a good conditional on trade ($\mu_{\leq b_t^*(t)(\alpha)}$) is greater (less) than the expected value of a traded good in a population of agents who choose the optimal threshold. Agents of type t_Γ choose the optimal threshold against as-if rational populations. Against a population of agents who choose a threshold that is too high (low), such as a population of types t, these types will choose a threshold that is too low (high), that is, $b_t^*\left(t_\Gamma\right)\left(\alpha\right)<b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)$ ($b_t^*\left(t_\Gamma\right)\left(\alpha\right)>b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)$).

2. Next we show that type $BR\left(t\right)$ eliminates type t. We have to show that type $BR\left(t\right)$ achieves a strictly higher payoff than type t in population $q\cdot BR\left(t\right)+\left(1-q\right)\cdot t$ for any $q\in\left(0,1\right)$, which implies that type $BR\left(t\right)$ eliminates t because the hybrid-replicator dynamics is payoff monotone when all agents have the same level of cursedness. Fix $\alpha\in\left[1,\frac{H}{L}\right]$. Assume that $b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)<b_t^*\left(t\right)\left(\alpha\right)$ ($b_t^*\left(BR\left(t\right)\right)\left(\alpha\right)>b_t^*\left(t\right)\left(\alpha\right)$). This implies that agents of type $BR\left(t\right)$ choose a lower (higher) thresholds than agents of type t. Recall that an agent of type $BR\left(t\right)$ chooses an optimal threshold against a population of agents of type t. Therefore, by a similar argument to that in the previous step, type $BR(t)$ chooses a threshold that is too high (low) against a mixed population $q\cdot BR\left(t\right)+\left(1-q\right)\cdot t$. But, this implies that the optimal threshold (for the given α) in the mixed population is strictly closer to the threshold used by type $BR\left(t\right)$ than to the threshold used by type t. Thus, agents of type $BR\left(t\right)$ make smaller mistakes in their thresholds for all values of α, and achieve a strictly higher payoff in the barter trade than that of the agents of type t in the mixed population. For a sufficiently low p, the same holds also in the game u_p.

3. The distance $\left\|t-BR\left(t\right)\right\|$ measures how much an agent of type t chooses wrong thresholds against a population of t-agents. The arguments in the proof of Prop. 1 imply

that $\|t - BR(t)\| = 0$ iff $t \in \Gamma$, and relatively simple adaptations of these arguments show that if $t \notin \Gamma_{\delta}$, then there is $\epsilon = O(\delta)$ such that $\|t - BR(t)\| > \epsilon$. The fact that $BR(t)$ is strictly between t and t_{Γ} (as shown in part (1) above) implies that

$$\|t_{\Gamma} - BR(t)\| = \|t_{\Gamma} - t\| - \|t - BR(t)\| < \|t_{\Gamma} - t\| - \epsilon.$$

A.9 Proof of Prop. 4 (Slow Drift within Γ)

Let $p^* \in (0,1)$ be sufficiently small such that$(1-r) \cdot u_p((0,I),t) < u_p(t,t)$ for each $t \in \Gamma$ and each $p \leq p^*$ (such $p^* > 0$ exists due to the fact that $u_0((0,I),t) = u_0(t,t)$, $r > 0$, and the Lipschitz continuity of the function $v(t,\eta)$. Without loss of generality we can assume that $\bar{p} < p^*$ (as, otherwise, we can choose a sufficiently large ϵ such that part (1) holds trivially, and the proof of part (2) is unaffected by taking a large ϵ).

1. Fix a sufficiently small $\bar{\delta} \ll \bar{p}, r$, such that $(1-r) \cdot u_{p^*}((0,I),t) < u_p(t,t)$ for each $t \in \Gamma_{\bar{\delta}}$ and each $p \leq p^*$. Let $t = (\chi,\psi) \in \Gamma_{\bar{\delta}}$. Observe that for each $\epsilon > 0$, if $t' = (\chi',\psi')$ is a type such that $|\chi' - \chi|, \|\psi - \psi'\| > \epsilon$, then both hybrid types (χ,ψ') and (χ',ψ) have thresholds that differ substantially (i.e., by more than $\mathcal{O}(\epsilon)$) from the optimal thresholds against a population of agents of type t. This implies that there is $\lambda = \mathcal{O}(\epsilon)$ such that $u_0(t,t) > u_0((\chi,\psi'),t) + \lambda$ and $u_0(t,t) > u_0((\chi',\psi),t) + \lambda$. This, in turn, implies (by the Lipschitz continuity of the payoff function of the additional activities v) that there is $\epsilon = \mathcal{O}(\bar{p})$ such that for each $p \in (0,\bar{p})$ and each type $t' = (\chi',\psi')$ satisfying $|\chi' - \chi|, \|\psi - \psi'\| > \epsilon$ the following inequalities hold: $u_p(t,t) > u_p((\chi,\psi'),t)$ and $u_p(t,t) > u_p((\chi',\psi),t)$. These inequalities imply, due to Part (4) of Lemma 2, that type t is asymptotically stable against type t'.
   Next, consider a sequence $(t_0 = t, t_1, ..., t_n = (I,0))$ that satisfies that each type t_{i+1} eliminates type t_i. The above argument implies that the distance between each two successive elements must be at most $\epsilon = \mathcal{O}(\bar{p})$. This, in turn, implies the following minimal bound on the length of the sequence:

   $$n \geq \frac{\|t\|}{\epsilon} = \|t\| \cdot \Omega\left(\frac{1}{\bar{p}}\right),$$

   and, in particular, the RHS converges to infinity as $\bar{p} \to 0$.

2. Fix $p \in (0,\bar{p})$ (where $\bar{p}$ is defined as above). Let $0 < \delta, \alpha \ll p$. Let $t = (\chi,\psi) \in \Gamma_{\delta} \setminus \{(0,I)\}$. Let $\lambda = \max(\alpha, \|\Gamma - t\|)$. Let $t' = (\chi',\psi)$ be a type with the same perception bias and a slightly lower level of cursedness, i.e., $\chi - \lambda < \chi' < \chi$. Consider

any mixed population $\eta_q = q \cdot t' + (1-q) \cdot t$. Observe that in any such population both types choose almost the same thresholds for each $\alpha \in \left[1, \frac{H}{L}\right]$:

$$b^*_{\eta_q}(t')(\alpha) = b^*_{\eta_q}(t)(\alpha) + O(\lambda),$$

and that these thresholds are almost optimal (because the thresholds of type t are almost optimal in a population of agents of type t). A similar "envelope-theorem" argument as in the proof of part 1(b) of Prop. 2 implies that the expected loss of type t' in the barter trade is bounded by $O(\lambda^2)$, i.e., $u_0(t',t) \geq u_0(t,t) - O(\lambda^2)$. On the other hand, type t' earns a strictly higher payoff of $\frac{\partial v}{\partial \chi}(\chi) \cdot O(\lambda)$ in the additional activities. This implies that $u_p(t',\eta) > u_p(t,\eta)$ for a sufficiently small λ, which implies that type t' eliminates type t (as the hybrid-replicator dynamics is payoff monotone if all agents have the same perception bias).

Remark 4. The proof above eliminates type t by a mutant with a slightly lower level of cursedness. The same argument can be adapted to show elimination by a mutant type with a slightly smaller endowment effect.

References

Alger, Ingela, & Weibull, Jörgen W. 2013. *Homo Moralis*: Preference Evolution under Incomplete Information and Assortative Matching. *Econometrica*, **81**(6), 2269–2302.

Apicella, Coren L., Azevedo, Eduardo M., Fowler, James H., & Christakis, Nicholas A. 2014. Evolutionary Origins of the Endowment Effect: Evidence from Hunter-Gatherers. *American Economic Review*, **104**(6), 1793–1805.

Bénabou, Roland, & Tirole, Jean. 2002. Self-Confidence and Personal Motivation. *The Quarterly Journal of Economics*, **117**(3), 871–915.

Bergstrom, Ted, & Bergstrom, Carl T. 1999. Does Mother Nature Punish Rotten Kids? *Journal of Bioeconomics*, **1**(1), 47–72.

Bernardo, Antonio E., & Welch, Ivo. 2001. On the Evolution of Overconfidence and Entrepreneurs. *Journal of Economics & Management Strategy*, **10**(3), 301–330.

Bokhari, Sheharyar, & Geltner, David. 2011. Loss Aversion and Anchoring in Commercial Real Estate Pricing: Empirical Evidence and Price Index Implications. *Real Estate Economics*, **39**(4), 635–670.

Bomze, Immanuel M. 1990. Dynamical Aspects of Evolutionary Stability. *Monatshefte für Mathematik*, **110**(3–4), 189–206.

Cesarini, David, Johannesson, Magnus, Magnusson, Patrik K. E., & Wallace, Björn. 2012. The Behavioral Genetics of Behavioral Anomalies. *Management Science*, **58**(1), 21–34.

Compte, Olivier, & Postlewaite, Andrew. 2004. Confidence-Enhanced Performance. *American Economic Review*, **94**(5), 1536–1557.

Cressman, Ross. 1997. Local Stability of Smooth Selection Dynamics for Normal Form Games. *Mathematical Social Sciences*, **34**(1), 1–19.

Dekel, Eddie, Ely, Jeffrey C., & Yilankaya, Okan. 2007. Evolution of Preferences. *Review of Economic Studies*, **74**(3), 685–704.

Ely, Jeffrey C. 2011. Kludged. *American Economic Journal: Microeconomics*, **3**(3), 210–231.

Eshel, Ilan, & Feldman, Marcus W. 1984. Initial Increase of New Mutants and Some Continuity Properties of ESS in Two-Locus Systems. *American Naturalist*, **124**(5), 631–640.

Eshel, Ilan, & Motro, Uzi. 1981. Kin Selection and Strong Evolutionary Stability of Mutual Help. *Theoretical Population Biology*, **19**(3), 420–433.

Eyster, Erik, & Rabin, Matthew. 2005. Cursed Equilibrium. *Econometrica*, **73**(5), 1623–1672.

Genesove, David, & Mayer, Christopher. 2001. Loss Aversion and Seller Behavior: Evidence from the Housing Market. *The Quarterly Journal of Economics*, **116**(4), 1233–1260.

Grosskopf, Brit, Bereby-Meyer, Yoella, & Bazerman, Max. 2007. On the Robustness of the Winner's Curse Phenomenon. *Theory and Decision*, **63**, 389–418.

Guth, Werner, & Yaari, Menahem. 1992. Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach. *Pages 23–34 of:* Witt, Ulrich (ed), *Explaining Process and Change: Approaches to Evolutionary Economics.* University of Michigan Press, Ann Arbor.

Hammerstein, Peter. 1996. Darwinian adaptation, population Genetics and the Streetcar Theory of Evolution. *Journal of Mathematical Biology*, **34**(5–6), 511–532.

Harrison, Glenn W., & List, John A. 2004. Field Experiments. *Journal of Economic Literature*, **42**(4), 1009–1055.

Haviland, W.A., Prins, H.E.L., Walrath, D., & McBride, B. 2007. *Anthropology: The Human Challenge*. 12th illustrated edn. Wadsworth/Thomson Learning.

Heifetz, Aviad, & Segev, Ella. 2004. The Evolutionary Role of Toughness in Bargaining. *Games and Economic Behavior*, **49**(1), 117–134.

Heifetz, Aviad, Shannon, Chris, & Spiegel, Yossi. 2007. The Dynamic Evolution of Preferences. *Economic Theory*, **32**, 251–286. 10.1007/s00199-006-0121-7.

Heller, Yuval. 2015. Three Steps Ahead. *Theoretical Economics*, **10**(1), 203–241.

Herold, Florian. 2012. Carrot or Stick? The Evolution of Reciprocal Preferences in a Haystack Model. *The American Economic Review*, **102**(2), 914–940.

Herold, Florian, & Netzer, Nick. 2015. *Second-best Probability Weighting*. Working Paper.

Herskovits, Melville J. 1952. *Economic Anthropology: A Study in Comparative Economics*. A. A. Knopf.

Huck, Steffen, Kirchsteiger, Georg, & Oechssler, Jorg. 2005. Learning to Like What You Have: Explaining the Endowment Effect. *The Economic Journal*, **115**(505), 689–702.

Jehiel, Philippe. 2005. Analogy-based expectation equilibrium. *Journal of Economic theory*, **123**(2), 81–104.

Jehiel, Philippe, & Koessler, Frédéric. 2008. Revisiting Games of Incomplete Information with Analogy-Based Expectations. *Games and Economic Behavior*, **62**(2), 533–557.

Johnson, Dominic D. P., & Fowler, James H. 2011. The Evolution of Overconfidence. *Nature*, **477**(7364), 317–320.

Kagel, John H., & Roth, Alvin E. (eds). 1997. *The Handbook of Experimental Economics*. Princeton University Press.

Kagel, John Henry, & Levin, Dan. 2002. *Common Value Auctions and the Winner's Curse*. Princeton University Press.

Kahneman, Daniel, & Lovallo, Dan. 1993. Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking. *Management Science*, **39**(1), 17–31.

Kahneman, Daniel, Knetsch, Jack L., & Thaler, Richard H. 1990. Experimental Tests of the Endowment Effect and the Coase Theorem. *Journal. of Political Economy*, **98**(6), 1325–1348.

Kahneman, Daniel, Knetsch, Jack L., & Thaler, Richard H. 1991. Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias. *Journal. of Economic Perspectives*, **5**(1), 193–206.

Karlin, Samuel. 1975. General Two-Locus Selection Models: Some Objectives, Results and Interpretations. *Theoretical Population Biology*, **7**(3), 364–398.

Knetsch, Jack, Tang, Fang-Fang, & Thaler, Richard. 2001. The Endowment Effect and Repeated Market Trials: Is the Vickrey Auction Demand Revealing? *Experimental Economics*, **4**, 257–269.

Liberman, Uri. 1988. External Stability and ESS: Criteria for Initial Increase of New Mutant Allele. *Journal of Mathematical Biology*, **26**(4), 477–485.

Lindsay, Luke. 2011. *Market Experience and willingness to trade: evidence from repeated markets with symmetric and asymmetric information*. Tech. rept. Working Paper Series, Department of Economics, University of Zurich.

List, John A. 2003. Does Market Experience Eliminate Market Anomalies? *The Quarterly Journal of Economics*, **118**(1), 41–71.

List, John A. 2004. Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace. *Econometrica*, **72**(2), 615–625.

Massey, Cade, & Thaler, Richard H. 2013. The Loser's Curse: Decision Making and Market Efficiency in the National Football League Draft. *Management Science*, **59**(7), 1479–1495.

Matessi, Carlo, & Di Pasquale, Cristina. 1996. Long-Term Evolution of Multilocus Traits. *Journal of Mathematical Biology*, **34**(5–6), 613–653.

Maynard Smith, John. 1971. What Use is Sex? *Journal of Theoretical Biology*, **30**(2), 319–335.

Nachbar, John H. 1990. Evolutionary Selection Dynamics in Games: Convergence and Limit Properties. *International Jjournal of Game Theory*, **19**(1), 59–89.

Norman, Thomas W. L. 2008. Dynamically Stable Sets in Infinite Strategy Spaces. *Games and Economic Behavior*, **62**(2), 610–627.

Oechssler, Jörg, & Riedel, Frank. 2002. On the Dynamic Foundation of Evolutionary Stability in Continuous Models. *Journal of Economic Theory*, **107**(2), 223–252.

Ok, Efe A, & Vega-Redondo, Fernando. 2001. On the evolution of individualistic preferences: An incomplete information scenario. *Journal of Economic Theory*, **97**(2), 231–254.

Polanyi, Karl. 1957. The Economy as Instituted Process. *Pages 243–270 of:* Polanyi, Karl, Arensberg, Conrad M., & Pearson., Harry W. (eds), *Trade and Market in the Early Empires: Economies in History and Theory*. The Free Press.

Robson, Arthur J., & Samuelson, Larry. 2009. The Evolution of Time Preference with Aggregate Uncertainty. *The American Economic Review*, **99**(5), 1925–1953.

Sahlins, Marshall David. 1972. *Stone Age Economics*. Aldine.

Sandholm, William H. 2001. Preference Evolution, Two-Speed Dynamics, and Rapid Social Change. *Review of Economic Dynamics*, **4**(3), 637–679.

Steiner, Jakub, & Stewart, Colin. 2016. Perceiving Prospects Properly. *American Economic Review*, **106**(7), 1601–31.

Taylor, Peter D., & Jonker, Leo B. 1978. Evolutionary Stable Strategies and Game Dynamics. *Mathematical Biosciences*, **40**(1), 145–156.

Thaler, Richard. 1980. Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior and Organization*, **1**(1), 39–60.

Thomas, Bernhard. 1985. On Evolutionarily Stable Sets. *Journal of Mathematical Biology*, **22**(1), 105–115.

Waldman, Michael. 1994. Systematic Errors and the Theory of Natural Selection. *The American Economic Review*, **84**(3), 482–497.

Weibull, Jörgen W. 1997. *Evolutionary Game Theory*. The MIT Press.