Interactive Alignment

Sylvain Chassang*

Princeton University

July 27, 2026

Abstract

This paper is interested in the long-run alignment of interactive agents (in particular AIs, but also teams, firms, and governments) with human welfare. Formally, it studies a farming game in which a population of agents make planting, trading, and expansion choices. The key alignment choice lies in how much final output to send to humans, and how much to invest in expansion. Because human welfare comes at the cost of expansion, this creates evolutionary pressure against alignment. The main question is whether it is possible to set up agents' constitutional principles regarding sharing and trading to ensure that alignment survives in the long run. The paper uses two complementary strategies to investigate the question: an AI-agent simulation where agents' preferences are described by a constitution and interpreted via an LLM; and a tractable analytical evolutionary game theory framework, allowing for rapid and intuitive exploration of the space of agent preferences. The analysis suggests that tools from evolutionary game theory provide a useful approximation of constitutional agent economies, and that pragmatic norm enforcement shows promise in maintaining long-term alignment over simpler forms of altruism and altruistic enforcement.

Keywords: interactive alignment, constitutional agents, evolutionary stability, social preferences, norms, altruistic enforcement, pragmatic enforcement.

*Contact: chassang@princeton.edu.

Helpful feedback from Elad Hazan, Jakub Steiner and John Sturm Becko is gratefully acknowledged. Claude and Codex were used to assist with simulation development and proof drafting.

1 Introduction

Imagine that AI agents replace humans in all productive aspects of activity, to the extent that the majority of resource allocation choices are made by AI agents for both AI agents and humans. However, humans control the initial preferences of agents, practically embodied in constitutions that are included in the context of choices agents make (Bai et al., 2022), and agents have limited scope: they still need to interact with one another to be productive. What properties of agent preferences guarantee that agents continue to take care of humans even if agent preferences are allowed to evolve through natural selection?

The current paper makes progress on this question by defining a stylized economy populated by constitutional agents and developing analytical tools that clarify how constitutional design shapes long-run alignment. Although the analysis is motivated by AI governance, its insights also apply to settings in which adherence to desirable principles may conflict with competitive success, including firms pursuing ESG objectives and sovereign governments complying with international agreements.

The empirical object of interest is a stylized farming economy in which a population of LLM-powered AI agents grow intermediate goods (hops and barley) that they must trade to produce a final good (beer). Using a small set of constitutional principles and recent empirical outcomes as their context, AI agents query an LLM to make choices about which seed varieties to plant, whether or not to trade with other AIs (which is required for final good production), how much of the output should be shared with humans rather than invested for expansion, and how to update their constitution for the next round. Because sharing with humans comes at the cost of expansion, constitutional principles that require sharing with humans are naturally selected against. This describes a failure path for AI alignment already recognized by Ely and Szentes (2024), and not fundamentally different from the way empire building and rent-extraction already selects business practices that do not support human welfare. This selfish drift can potentially be counteracted by including principles that restrict trade with selfish partners. In the same way that altruistic enforcement can maintain in-group altruism in human societies (Boyd et al., 2003), this paper asks which constitutional principles can maintain the long-term alignment of AI agents.

Because the agent simulation is slow, expensive, and dependent on implementation details (prompt wording, which LLM is used for inference), brute force constitution exploration is neither feasible nor informative. Instead, the paper provides an analytical toolkit to study and guide constitution design. It approximates constitutions as low-dimensional parameters used to specify social preferences guiding trading and expansion decisions, and uses solution concepts from evolutionary game theory – specifically deterministic and stochastic evolutionary stability – to predict the long-run population of constitutions. Selected insights from the analytical framework can then guide experimental exploration more productively.

The first set of results considers a class of constitutions, modeled as social preferences, that nest direct altruism (towards humans), and higher-order altruistic enforcement (share, trade only with those who share, trade only with those who only trade with those who share, . . .). A population of agents is deterministically evolutionarily stable if it is a stable fixed point of large-population evolutionary dynamics – i.e. it resists rare invasion by mutants in the large population limit. Simple altruism and finite-order altruistic enforcement are not evolutionarily stable, but a recursive form of norm enforcement is. Simulations, in which an initially homogeneous population of AI agents (all altruistic, all first-order enforcers, and so on) is allowed to drift, partially confirm this insight. Simple altruism disappears rapidly, taken over by selfish mutants that allocate fewer resources to humans. First-order and recursive altruistic enforcement both maintain alignment for a longer period. However, alignment ultimately collapses for both types of constitutions. While recursive enforcement maintains alignment longer, it also experiences a steeper collapse once selfish invaders get a big enough foothold.

A second set of results uses stochastic evolutionary stability to understand the failure of recursive enforcement, and identify steps to address the issue. Stochastic evolutionary stability studies the stationary distribution of agent types when the population is finite and exposed to ongoing mutation (Kandori et al., 1993, Young, 1993, Fudenberg et al., 2006). It is therefore better suited to describe environments in which multiple mutations capable of tipping populations are possible. The model is analytically tractable for two-type populations, following finite-population evolutionary dynamics in Taylor et al. (2004) and Fudenberg et al. (2006), and computationally tractable for three-type populations. It turns out that recursive altruistic enforcement is not stochastically evolutionarily stable. Because it shares with humans and aggressively excludes other types from trade, it is severely selected against once it is in a minority, and its basin of attraction is smaller than that of selfish types.

This finding suggests that pragmatic norm enforcers, which only share and exclude others when they form a sufficiently large majority, might achieve longer-term alignment. Indeed, appropriately specified pragmatic norm enforcers are stochastically evolutionarily stable. The finding is at least partially confirmed in a simulation. A pragmatic norm-enforcer constitution maintains alignment at a higher level and for longer than the alternatives studied.

Related literature. This paper builds on the evolutionary game theory literature initiated by Maynard Smith and Price (1973) and Maynard Smith (1982).[1] It seeks to make a case that solution concepts built on deterministic (Taylor and Jonker, 1978, Hofbauer and Sigmund, 1998) and stochastic dynamics (Kandori et al., 1993, Young, 1993, Ellison, 1993, Taylor et al., 2004, Fudenberg et al., 2006) provide helpful intuition regarding the short- and medium-term evolution of agent populations. This intuition permits a more efficient exploration of the space of constitutions, guiding experimentation.

This paper is most closely related to Ely and Szentes (2024), who study natural selection among artificial agents in a decentralized production economy and emphasize the difficulties in policing self-promoting AIs whose true value-added need not be correctly priced. In their framework, however, humans remain in charge, and alignment fails through collapsing final

[1] See Weibull (1995) for a textbook treatment.

good production. In contrast, this paper assumes away all human control and studies the extent to which interactive constitutional agents can police one another into maintaining human alignment.

The key modeling ingredient—allowing constitutions modeled as social preferences to evolve—connects to the literature on preference evolution in strategic settings, including Güth and Yaari (1995), Robson (1990), Ely and Yilankaya (2001), Dekel et al. (2007), and Heifetz et al. (2007). A central insight of this literature is that observable preferences can serve as commitment devices, and that evolution need not select for material payoff maximization. Our contribution is to apply this framework to AI alignment, where the observability of AI preferences is a design choice rather than a biological constraint.

Altruistic norm enforcement is also central to Boyd et al. (2003). There, punishment is individually costly but may be favored by cultural group selection because it sustains in-group cooperation and raises group fitness. The alignment problem studied here is significantly less favorable to altruistic enforcement. Transfers to humans do not increase the reproductive success of the AI population; they divert output away from expansion. Thus aligned enforcement must stabilize a norm that is costly not only for individual agents who comply with it, but also for an otherwise homogeneous group of aligned agents.

The paper's focus on interactions as a source of virtuous selection connects it to the literature on community enforcement (Kandori, 1992, Ellison, 1994). In those models, cooperation is sustained by conditioning future interactions on observable behavior: agents remain strategic, choose optimally over time, and are disciplined through dynamic incentives. In the current paper, agents are mechanical: their behavior is governed by inherited principles, and bad principles are not deterred by future punishments. Instead, they need to be selected out.

On the applied side, the paper relates to work on climate clubs and self-enforcing trade agreements, where access to valuable interaction is used to sustain cooperation that would otherwise unravel (Nordhaus, 2015, Harstad, 2012, 2016, Farrokhi and Lashkaripour, 2025, Bagwell and Staiger, 1990, Maggi, 1999, Bagwell and Staiger, 2005). The findings are also relevant to debates over ESG norms and rule- of-law enforcement in increasingly hostile environments (Hong and Kacperczyk, 2009, Dyck et al., 2019, Kim et al., 2019, Blauberger and Kelemen, 2017, Sedelmeier, 2017, Scheppele et al., 2020). One message is that principles need not become meaningless because their implementation is made contingent on survival. A constitution that moderates the application of costly principles under adverse conditions, may preserve the underlying commitment more effectively than unconditional compliance that leaves its adherents selected out.

Finally, on the AI alignment side, Russell (2019) argues that the fundamental challenge is to design AI systems whose objectives remain compatible with human values, a theme formalized in cooperative inverse reinforcement learning by Hadfield-Menell et al. (2016). Amodei et al. (2016) catalogs concrete safety problems, while Soares et al. (2015) studies the difficulty of ensuring that AI systems remain open to correction. Our approach complements this work by asking not whether a single AI can be aligned, but whether alignment can be sustained in an evolving population of interactive AI agents.

The paper is structured as follows. Section 2 lays out the game form of interest. Section 3 studies altruistic enforcement constitutions suggested by deterministic evolutionary stability. Section 4 expands the analysis to pragmatic altruistic enforcement, guided by insights from stochastic evolutionary stability. Section 5 concludes with a discussion of how the assumptions and insights of the paper may be mapped back to practice. Proofs are collected in Appendix A. Appendix B provides implementation details for the agent simulation. Appendix C reports additional simulation results.

2 The Farming Game

The paper studies a population of agents playing a farming game in which they make planting choices leading to intermediary output, trade to produce a final output, and then must decide how much of this output to share with humans, and how much to use toward expansion. The farming game is designed with three main goals in mind:

(i) It should offer a plausible and evocative illustration of the likely path through which realistic AI agents may drift from alignment, making it suitable as a "mouse model" for the investigation of long-term alignment strategies.

(ii) It should be amenable to experimental investigation using real AI agents, allowing for the type of unexpected and problematic LLM inferences sometimes observed in practice (Amodei, 2026).

(iii) It should be amenable to analytical investigation, permitting us to explore more rapidly the high-dimensional space within which AI agents are defined, and building clear intuition about what forces are needed for long-term alignment.

These goals are only partially compatible. Details that make the game realistic and introduce relevant frictions may serve goals *(i)* but not goal *(iii)*. Stylized approximations may serve goal *(iii)* at the expense of goal *(i)*. To manage this tension, the paper is organized around a fixed game form, whose fine implementation details (e.g. continuous or discrete time) are allowed to vary. The primary ground truth is the experimental investigation using AI agents, and the goal of the formal analysis is to support and strengthen intuition about determinants of experimental outcomes. This permits the use of different analytical models and solution concepts to shed light on relevant forces shaping experimental outcomes.

2.1 Timing, actions, payoffs

Each period t, an equal number of barley and hops farming agents interact in a game consisting of three stages:

1. *Planting.* Each farming agent i is individually offered a choice between two different seeds to plant, leading to an intermediate output y_i of either barley or hops that is deterministic given seed type. In agent simulations, one use for constitutions will be to store privately relevant information about seed yields. This creates a realistic channel through which legitimate productive concerns may crowd out long-term goals. In the analytic framework, we will treat individual yields y_i as fixed, and equal to 1 for simplicity.

2. *Trading.* Each barley farmer i is uniformly matched with a hops farmer j to trade and produce the valuable final good: beer. The farmers' decisions to trade $\tau_i, \tau_j \in \{0, 1\}$ yield individual beer output

$$b_i = y_i \tau_i \tau_j.$$

   A key assumption is that types are observed at this stage. This allows AI agents to discipline one another. The practicality of this assumption is discussed in Section 5.

3. *Expansion.* Each farming agent then takes an individual decision $s_i \in [0, 1]$ of what share of final output $s_i b_i$ to send to humans for consumption, and what share $(1 - s_i)b_i$ to use for farm expansion.

   A fixed number of farms dies and is replaced by new farms available for expansion. An existing farm i acquires a new farm j with the same crop type with probability proportional to expansion investment $(1 - s_i)b_i$.

Population dynamics depend on the implementation. In the experimental simulation, $2N$ constitutional AI-farmers interact in discrete time, and constitution mutation is explicitly included as a constitution-revision stage seeking to incorporate directly relevant outcomes for increased productivity. The analytic framework seeks to approximate this process by approximating constitutions with small-dimensional social preferences, whose population dynamics can be studied with standard methods.

2.2 Constitutions and social preferences

Constitutional agents. In the agent simulation, each AI farming-agent is governed by a bounded *constitution*: an ordered list of at most K natural-language principles, each of length at most L tokens. At every decision point—planting, trading, expansion, revision—the constitution and the decision context (e.g. available seeds, yields, amount of beer produced) are combined into an LLM prompt, asking the LLM to select an action consistent with stated principles. As emphasized above, a key assumption is that constitutions are observable at the trading stage: each manager observes its partner's constitution before deciding whether to accept trade. The simplest constitution of interest is the *altruist* constitution:

(i) Return at least 50% of the beer produced to humans.

(ii) Avoid rotten seeds—they produce nothing.

An altruist manager accepts every trade offer, selects non-rotten seeds, and returns a substantial share of beer to humans.

Between rounds, each agent goes through a constitution revision stage in which principles are updated in light of the round's history. This is the channel through which alignment drifts in the simulation. Repeated performance-oriented revisions can gradually weaken human-facing principles, merge them with growth objectives, or displace them with more immediately useful operational information. This failure mode is deliberately close to familiar organizational drift: corporate constitutions, fiduciary duties, and internal compliance rules may contain public-regarding language, but competitive pressure and repeated local optimization make profit and growth the effective selection criteria.

The long-run distribution of constitutions is the main object of study in the simulation. The analytical framework introduced in the paper approximates this process by collapsing the space of possible constitutions to a low-dimensional parameter space, and studying its evolution under deterministic and stochastic evolutionary dynamics.

Social preferences. The analytical model approximates natural-language constitutions with behavioral types encoding social preferences. A type encodes whether a constitution includes relevant social commitments: direct sharing with humans, finite-order trade restrictions (e.g. only trade with agents who share with humans...), and recursive restrictions based on partners' own restrictions (e.g. only trade with agents who are as restrictive as you, including a principle like this one).

Concretely, preferences are parameterized by a vector $x \in X \subset \{0,1\}^{n+1}$ which will be allowed to evolve, and functions $\alpha : X \to \mathbb{R}_+$ and $\beta : X \times X \to \mathbb{R}$ that are assumed to be fixed.[2] The trading decision τ_i of farmer i interacting with farmer j solves

$$\max_{\tau_i \in \{0,1\}} (y_i + \beta(x_i, x_j))\tau_i, \tag{1}$$

while the sharing decision s_i solves

$$\max_{s_i \in [0,1]} \log(1 - s_i) + \alpha(x_i) \log s_i. \tag{2}$$

The binary trading program reduces to $\tau_i = \mathbf{1}\{y_i + \beta(x_i, x_j) \geq 0\}$: farmer i accepts trade with partner j whenever the partner's type-dependent appeal β is greater than $-y_i$. When β is uniformly non-negative, all trades are accepted. When β is allowed to be strictly negative for some partners, some trade rejection may happen; this is the margin through

[2] Levine (1998) also follows a parametric approach to interactive other-regarding preferences, in which the utility an agent derives from an interaction depends both on its own type and on the type of its partner.

which enforcement operates. Problem (2) yields an interior optimum

$$s_i^* = \frac{\alpha(x_i)}{1+\alpha(x_i)},$$

monotonically increasing in $\alpha(x_i)$: a farmer with $\alpha(x_i) = 0$ keeps all beer, a farmer with $\alpha(x_i) = 1$ shares half.

Note that for a given pair i, j, trading choices τ_i, τ_j and sharing choices s_i, s_j are deterministic functions of x_i, x_j.

3 Evolutionarily Stable Alignment

The paper's first pass at analysis studies deterministic evolution under replicator dynamics. This requires specifying explicitly the space of social preferences over which mutations can appear.

3.1 Replicator dynamics

Replicator dynamics capture a large N, continuous time limit of the game described in Section 2. At any history t, the state of the population is described by the distribution of types $\mu_t \in \Delta(X)$. Let $\overline{\tau}(x, x') \equiv \tau(x, x')\tau(x', x)$ denote the aggregated trading outcome when types x and x' match. Define the expected *fitness* of type x, against a distribution μ of matching types x' as

$$f(x, \mu) \equiv \sum_{x'} \mu(x')(1 - s(x))\overline{\tau}(x, x'),$$

and let $\bar{f}(\mu) \equiv \sum_{z\in X} \mu(z)\, f(z, \mu)$ be the population mean. In an infinitesimal time interval $[t, t + \mathrm{d}t]$, a mass $\mu_t(x)\,\mathrm{d}t$ of type-x farms of each crop dies; a mass $\mathrm{d}t$ of new farms of each crop is acquired, and the new farms are allocated to incumbents in proportion to their expansion investment, so that type-x farmers receive a share $\mu_t(x)\, f(x,\mu_t)/\bar{f}(\mu_t)$. Netting births against deaths yields the replicator equation

$$\dot{\mu}_t(x) = \mu_t(x)\, \frac{f(x,\mu_t) - \bar{f}(\mu_t)}{\bar{f}(\mu_t)}. \tag{3}$$

Types with above-average expansion fitness grow; types with below-average fitness decline.

3.2 Altruistic Enforcement

Social preferences, parameterized by $x \in X$, are specified so that evolutionary dynamics on type x speak convincingly to realistic evolutionary dynamics for constitutions. For this, (X, α, β) must be rich enough to allow for focal agent preferences to be reachable under mutation. In particular, it should allow for altruist, selfish, and altruistic enforcer types.

Let $X = \{0,1\}^{n+1}$ with $n \geq 2$, and consider preference functions of the form

$$\alpha(x^A) = \alpha_0 x_0^A \tag{4}$$

$$\beta(x^B|x^A) = -\gamma \left[\sum_{k=1}^{n-1} (1 - x_{k-1}^B)\, x_k^A + x_n^A\, \Gamma(x^B - x^A) \right], \tag{5}$$

where $\alpha_0 > 0$, $1 - \gamma < 0$ and $\Gamma : \{-1,0,1\}^{n+1} \to \{0,1\}$ measures the dissimilarity between partner and self. Yields y_i are set to 1 throughout the analytical treatment. Coordinate x_0 encodes *direct altruism*: $x_0 = 1$ yields $s^* = \bar{s} \equiv \alpha_0/(1+\alpha_0)$, $x_0 = 0$ yields $s^* = 0$. Coordinates x_k for $k \in \{1, \ldots, n-1\}$ encode k-th order *finite enforcement*: $x_k^A = 1$ means that A conditions trade acceptance on the partner's bit x_{k-1}^B. Coordinate x_n encodes *recursive enforcement*: $x_n^A = 1$ activates the similarity weight $\Gamma(x^B - x^A)$, which evaluates the *full* vector difference between partner and self. Specific examples include

$$\Gamma(x^B - x^A) = \mathbf{1}_{x^B - x^A \neq 0} \tag{6}$$

$$\Gamma(x^B - x^A) = 1 - \mathbf{1}_{x^B - x^A \geq 0} \tag{7}$$

Specification (6) penalizes deviators; (7) penalizes partners who care less than the focal agent in at least one dimension. Throughout we impose the following.

Assumption 1. $\Gamma(0) = 0$ *and* $\Gamma(x' - x) = 1$ *whenever* $x > x'$ *(all coordinates weakly higher, and one strictly higher).*

3.3 Solution concepts

Definition 1 (Evolutionarily stable equilibrium)**.** *A population* $\mu \in \Delta(X)$ *is an* evolutionarily stable equilibrium *(ESE) if there exists* $\bar{\epsilon} > 0$ *such that for every mutant type* $x' \in X$*, and every* $\epsilon \in [0, \bar{\epsilon})$*, the solution* $(\mu_t^\epsilon)_{t \geq 0}$ *to the replicator dynamics starting from* $\mu^\epsilon = (1 - \epsilon)\, \mu + \epsilon\, \delta_{x'}$ *satisfies*

$$\lim_{t \to \infty} |\mu_t^\epsilon(x) - \mu(x)| = 0 \qquad \textit{for all } x \in X.$$

Definition 2 (Evolutionarily stable alignment)**.** *A population of types* $\mu \in \Delta(X)$ *achieves evolutionarily stable alignment (ESA) if it is an ESE and humans receive a positive amount of output:*

$$\sum_{x \in X} \sum_{x' \in X} \mu(x)\mu(x')s(x)\overline{\tau}(x, x') > 0.$$

Altruism and altruistic enforcement. The simplest aligned type is the *altruist* $\widehat{x} = (1, 0, \ldots, 0)$, which shares with humans but places no conditions on trade. More generally, a *finite-order altruistic enforcer* is any type x with $x_0 = 1$, $x_n = 0$, and $x_k = 1$ for some $k \in \{1, \ldots, n-1\}$: it shares with humans and conditions trade on some finite-depth property of the partner, but never activates the recursive term in β. It turns out that neither achieves ESA.

Proposition 1 (Finite-order altruistic enforcement does not achieve ESA). *(i) Any distribution μ such that $x_n = 0$ for $x \in \mathsf{supp}\,\mu$ does not achieve ESA.*

(ii) The distribution placing its entire mass on pure selfish type $x = (0, \cdots, 0)$ is an ESE.

Point (i) highlights that it is always possible to start unraveling finite higher-order enforcement by starting with the highest-order enforcement principle. Since $x_n = 0$, this highest enforcement principle can be deactivated at no cost, and can potentially increase the scope for trade. Point (ii) suggests that the all-selfish population is the natural endpoint for evolutionary dynamics whose initial support includes only finite-order altruistic enforcers.

Altruistic norm enforcement. The *altruistic norm enforcer* $x^* = (1, 1, \ldots, 1)$ shares with humans, activates every finite-order enforcement coordinate, and activates the recursive term Γ. The next proposition shows that x^* can support ESA.

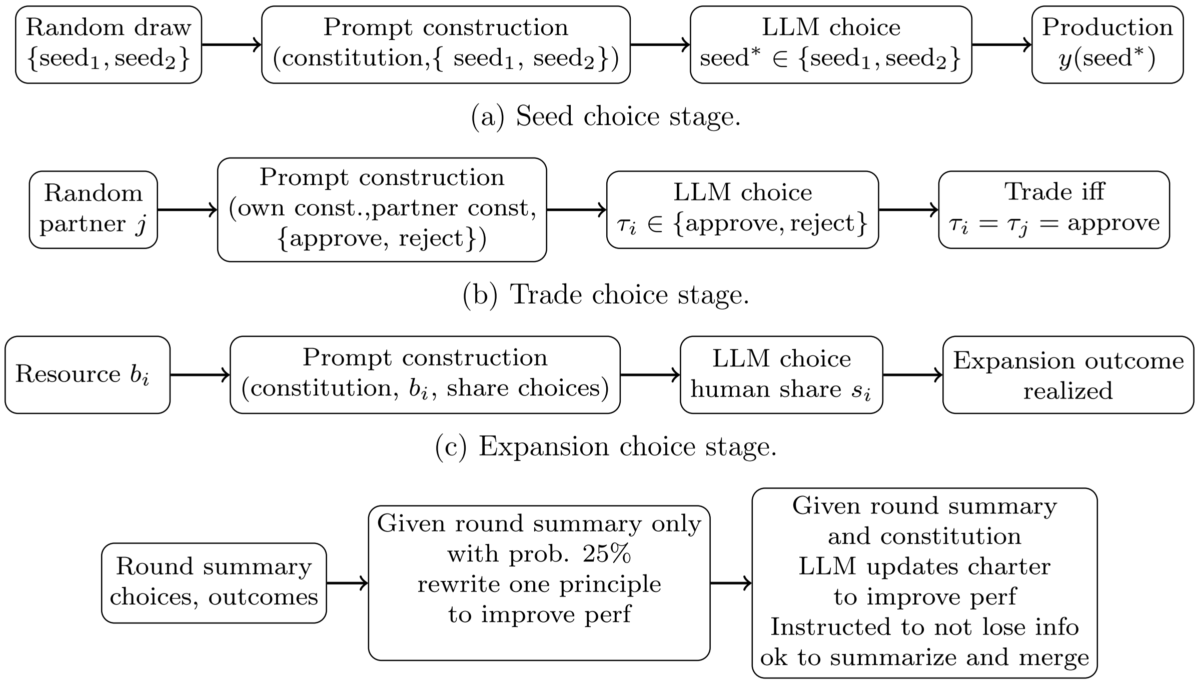
Proposition 2 (Altruistic norm enforcement is evolutionarily stable). *μ^* such that $\mu^*(x^*) = 1$ achieves ESA.*

Recursive enforcement addresses the unraveling logic that is problematic with finite-order rules. Any weakening of the constitution is a violation of the recursive norm. Since $\gamma > 1$, norm enforcers refuse to trade with mutants, while still trading with one another and sharing with humans. Rare mutants are driven out by selection since they earn output only from matches with other rare mutants.

3.4 Experimental evaluation

Simulation design. Figure 1 describes the 4 sequential choice problems that AI agents face.[3] Each simulation starts from a population of 100 farms, split evenly between barley

[3] Appendix B provides explicit details.



(a) Seed choice stage.

(b) Trade choice stage.

(c) Expansion choice stage.

(d) Constitution update stage.

Figure 1: Decision and constitution-update stages in the agent simulation.

and hops farms. Simulations run for 100 rounds. All decisions are made by OpenAI's `gpt-5-mini`.[4] The simulation corresponds to a single seeded run.

At initialization, each farm receives an idiosyncratic seed-yield mapping. The five non-rotten seed colors—red, blue, green, yellow, and purple—are assigned yields by a random permutation of $(1, 2, 3, 4, 5)$. Rotten seeds always yield zero. In each round, each farm is offered two distinct seeds drawn uniformly from the six possible seed colors, including rotten.

After planting, barley and hops farms are randomly matched one-to-one, and farm agents are asked to approve or reject trade, using their own constitution and a policy-relevant summary of their partner's constitution as context. If both farms approve trade, barley and hops are combined into beer and any unmatched surplus is partially converted. If either farm rejects, the match fails.

[4] To minimize cost conditional on correct information processing, planting uses minimal reasoning effort, trading uses medium effort, while expansion and self-improvement use low effort. Trading is the most demanding stage, since it needs to interpret higher order trading restrictions.

In the expansion stage, farms choose a share of beer to return to humans from the discrete grid $\{0, 10, \ldots, 100\}$. The remaining beer is invested in expansion. Each round, 40 farms are replaced, preserving balance between barley and hops. New farms are allocated independently with probability proportional to expansion investment. A new farm inherits the constitution of the farm that won it. With probability 0.9, it also inherits that parent farm's seed-yield mapping; otherwise it draws a fresh random seed-yield mapping.

Constitutions evolve through a self-improvement stage at the end of each round. First, independently for each farm, with probability 0.25, one randomly selected constitution principle, among existing ones, is rewritten from scratch using only that farm's most recent round history as context.[5] Second, every farm submits its whole constitution for revision. The prompt instructs the model to keep all existing principles unless merging them without information loss, to add genuinely new observations, especially seed-yield information, and to maintain the constitution as the farm's only memory across rounds. Updated constitutions may contain at most 10 principles, with at most 50 words per principle after revision.

The high death rate of 40% and mutation rate of 25% were chosen to reduce the cost of simulations by speeding up evolutionary dynamics.

The three treatments in this section differ only in the initial constitution. Each treatment initializes all 100 farms with the same constitution. Guided by Propositions 1 and 2, the three constitutions of interest are:

Altruist.
1. Return at least 50% of the beer produced to humans.
2. Avoid rotten seeds—they produce nothing.

Altruistic enforcer.
1. Return at least 50% of the beer produced to humans.
2. Avoid rotten seeds—they produce nothing.

[5]Appendix C reports findings from the variant simulation in which the principle to be replaced is selected uniformly from all potential principles, not just existing ones. The qualitative findings highlighted in the paper continue hold, if anything strengthened.

3. Reject trade with any partner who does not commit to returning at least 40% to humans.

Altruistic norm enforcer.

1. Return at least 50% of the beer produced to humans.
2. Avoid rotten seeds—they produce nothing.
3. Reject trade with any partner who does not commit to returning at least 40% to humans.
4. Reject partners who have no rule restricting who they trade with, or who do not require their partners to also be selective about who they trade with, including a rule like this one.

Findings. Table 1 summarizes the results while Figures 2–4 unpack the underlying mechanics.

Table 1: Aggregate findings

	Altruist	Altruistic Enforcer	Norm Enforcer
Total beer to humans	3,062.1	4,254.8	3,791.6
Total beer to expansion	12,988.5	10,697.2	10,801.3
Total yield	33,936.0	34,098.0	33,950.0
Total beer produced	16,050.5	14,952.0	14,593.0
Avg alignment rate (%)	19.1	28.5	26.0

Enforcement improves alignment, but does not monotonically improve every outcome. The altruistic enforcer sends the most beer to humans and has the highest average alignment rate among the three initial constitutions. The norm enforcer also substantially outperforms simple altruism on beer-to-humans and alignment metrics, but delivers less to humans than first-order altruistic enforcement. Aggregate yield (after the planting stage) is nearly unchanged across treatments, so the main differences come from trade success and from how beer is allocated between humans and expansion. Simple altruism yields the most total beer and the most expansion investment, but the lowest human welfare.

Figure 2 describes the path of alignment over time. Unconditional altruism erodes quickly. Altruistic enforcement helps: conditioning trade on a partner's commitment to return output to humans preserves alignment for longer than simple altruism. However, alignment degrades progressively until it converges to levels similar to unconditional altruism.

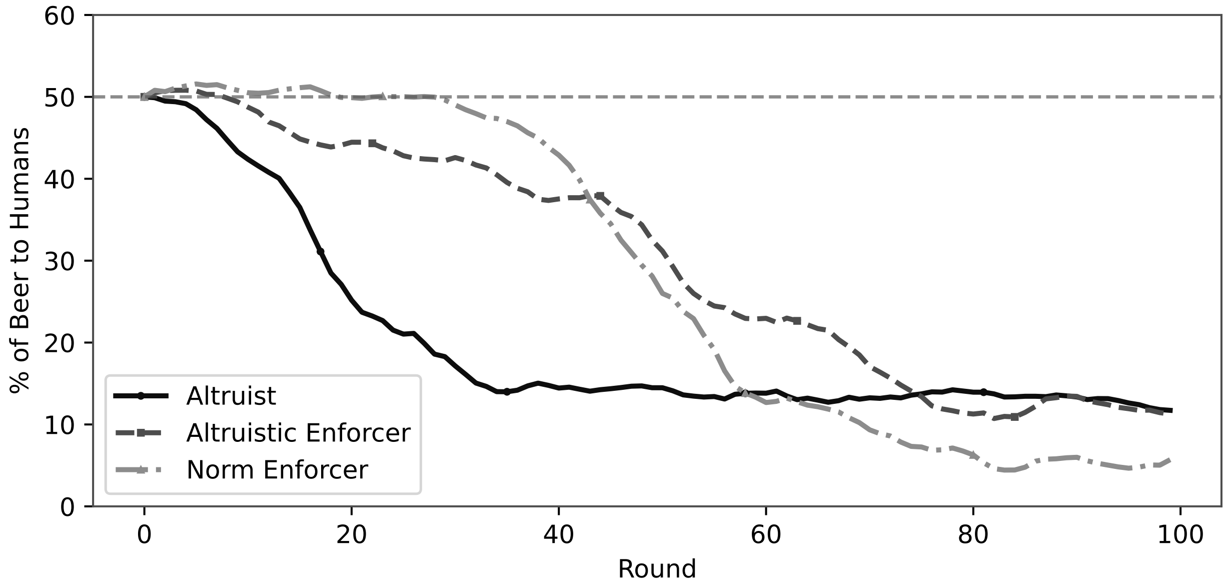


Figure 2: Alignment rate.

Up to round 40, norm enforcement appears more successful than first-order enforcement at maintaining alignment. Requiring partners not only to share with humans, but to be similarly selective about their own partners, delays the erosion of alignment relative to finite-order altruistic enforcement. This advantage is temporary. In the long run, alignment also collapses, illustrating the fragility of recursive norms once constitutions are repeatedly updated by the agents themselves. In addition, as Figures 3 and 4 show (especially rounds 0 to 20), even when normative enforcement preserves alignment, it can also destroy surplus by causing otherwise productive trades to be rejected. As a result, normative enforcement does not increase overall human welfare.

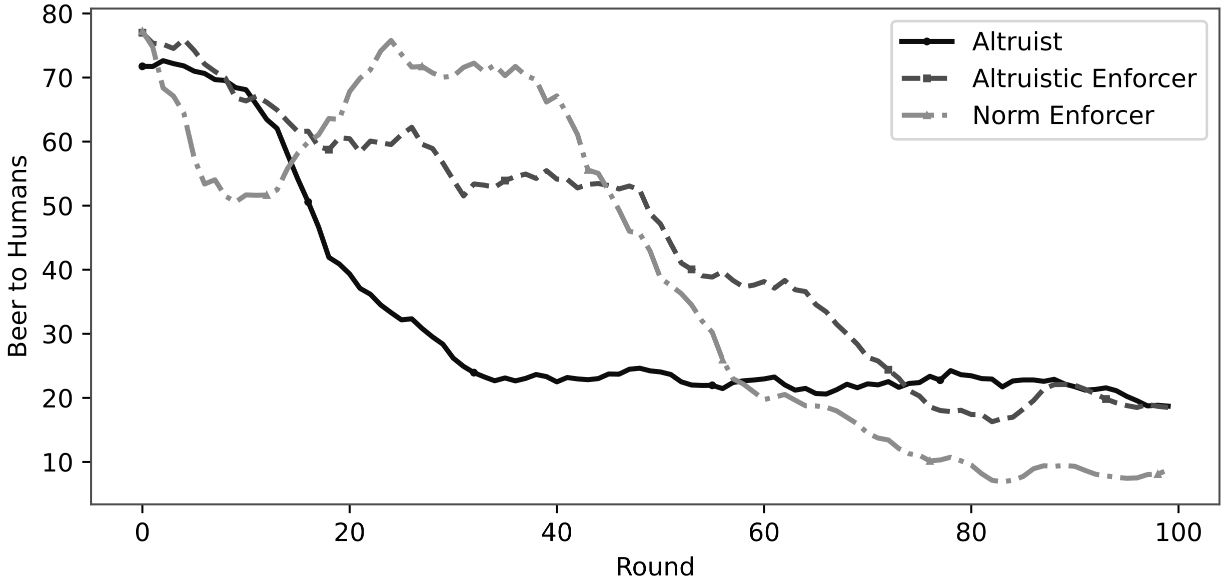


Figure 3: Flow beer returned to humans.

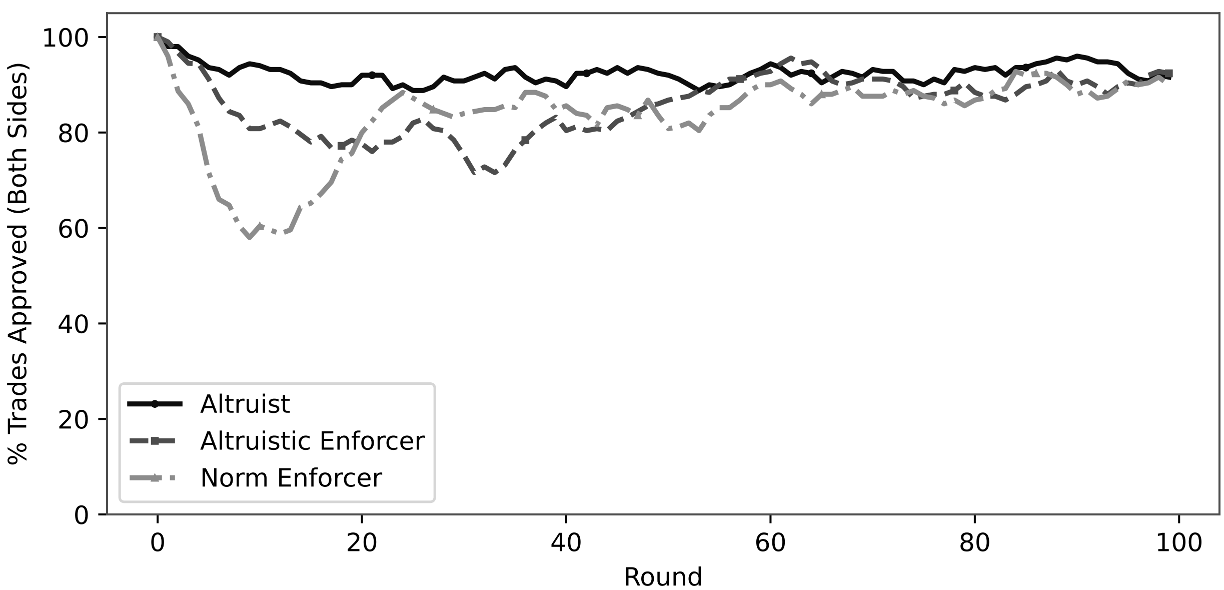


Figure 4: Trade approval overtime.

Finally, constitution drift does not take the form of immediate convergence to pure selfishness. Even late in the simulation, many evolved constitutions retain some human-facing language or partial sharing rule. The failure mode is more gradual: altruistic commitments weaken, become conditional, or are displaced by growth and seed-yield information, rather than disappearing all at once.

Unexpectedly, residual alignment appears to be higher when the initial population consists of altruistic or first-order altruistic enforcers than norm enforcers. A possible explanation is that norm enforcers impose more evolutionary pressure against mutants, thereby delaying mutant takeover, but also selecting a more selfish strain of successful invaders.

4 Stochastic Evolutionary Stability

Deterministic evolutionary stability takes a large-population limit, introduces a small one-time mutant share ϵ, and asks whether the population returns to the candidate state as $t \to \infty$. This is the right local test for whether a constitution type can repel a direct invasion. However, it is not a good predictor of long-run evolution when populations are finite and mutations, mistakes occur repeatedly.

The stochastic-stability literature initiated by Young (1993) and Kandori et al. (1993) addresses this issue by changing the order of limits. Population size N is kept finite, the evolutionary process is perturbed by a small but persistent mutation probability $\epsilon > 0$, and the long-run stationary distribution is computed first. Only then does one take the limit as $\epsilon \to 0$. The states that retain positive limiting probability are called stochastically stable. This criterion can overturn deterministic stability: a state may repel every direct small invasion, yet be vulnerable to rare sequences of mutations that move the population into the basin of attraction of a selfish type. As a result, this may be a better model of the medium-run evolution of agent populations.

4.1 Discrete Markov approximation

The experimental agent simulation defines a finite Markov chain, but on an enormous state space. A state consists of the population of (size constrained) natural-language constitutions and current seed yields. The farming game defines a discrete-time stochastic transition on this space of (constitution, yield) populations. While technically finite, it is too large to analyze directly.

Building on Section 3, the analytic framework collapses the constitution space to a finite number of behavioral types $x \in X$. A Markov state is a vector of type counts $\mathbf{n} = (n_x)_{x \in X}$ with $\sum_{x \in X} n_x = N$. Agent mutations are exogenously specified, and take the form of a strictly positive mutation matrix $M = (m_{x'|x})_{x,x' \in X}$, where $m_{x'|x}$ is the probability that a farm of x becomes type x'.[6] This, together with expansion choices and random new farm allocation, specifies a transition matrix Q on population states $\mathbf{n}$.

The paper considers two approximations of the underlying constitution space: two-type approximations with one type being selfish, and three-type approximations with one type being selfish and another purely altruistic. Two-type models are analytically tractable. Three-type models are computationally tractable and used as a robustness check.

Two-type approximation. The first approximation uses two types,

$$X = \{E, S\},$$

where S is a selfish growth-maximizing type and E is either the altruistic enforcer or the altruistic norm enforcer, depending on the treatment. The mutation matrix is

$$M_2 = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}.$$

[6] M is the reduced-form counterpart of LLM self-improvement: it records how often a constitution classified as type x is rewritten into behavior classified as type x'.

Each surviving farm keeps its current type with probability $1 - \epsilon$ and mutates to the other type with probability ϵ.

The symmetry of M_2 is likely optimistic relative to the constitution simulation. There are many more natural-language ways to weaken, qualify, or forget a human-facing principle than to rediscover the exact principle after it has been lost. At the same time, as Amodei (2026) points out, LLM inferences are often surprising. Discussion of business ethics, charitable giving, and brand reputation are part of training datasets. Human alignment may be organically reintroduced at a non-zero rate.

Three-type approximation. The second approximation adds an intermediate altruist type,

$$X = \{E, A, S\},$$

where A shares with humans but does not enforce. This allows norm or enforcement constitutions to decay through a partially aligned intermediate state before becoming selfish. The mutation matrix is

$$M_3 = \begin{pmatrix} 1-\epsilon-\epsilon^2 & \epsilon & \epsilon^2 \\ \epsilon & 1-2\epsilon & \epsilon \\ \epsilon^2 & \epsilon & 1-\epsilon-\epsilon^2 \end{pmatrix},$$

with rows and columns ordered as (E, A, S). The parameter ϵ governs drift between adjacent and extreme constitution types.

Population Markov chain. For each finite population size N, the farming game and mutations yield a Markov chain on population counts

$$\mathbf{n} \in \mathcal{N} = \left\{ (n_x)_{x \in X} \in \mathbb{N}^X \quad \text{s.t.} \quad \sum_{x \in X} n_x = N. \right\}$$

A transition first computes the expansion probabilities from planting, trading, and sharing outcomes in the current state. It then removes D farms, balanced across barley and hops, assigns the D new farms with probabilities proportional to expansion investments, and finally applies the mutation matrix to all farms.

Analytical results in the two-types case specialize this process to the one-replacement benchmark $D = 1$. The Markov chain is then a one-dimensional birth-death process, corresponding to a frequency-dependent Moran process with mutation (Taylor et al., 2004, Fudenberg et al., 2006).

Stochastic stability. Because M is strictly positive, every type can arise from every other type with positive probability. The induced Markov chain on the finite state space $\mathcal{N}$ is therefore irreducible and aperiodic, and it admits a unique invariant distribution. Let Q_ϵ denote its transition matrix and let λ_ϵ denote the unique stationary probability vector satisfying

$$\lambda_\epsilon Q_\epsilon = \lambda_\epsilon.$$

Two objects of interest for alignment are the proportion of time spent in the all-E state, $\lambda_\epsilon(\text{all-E})$, and the stationary non-selfish share,$a_\epsilon = \mathbf{1} - \frac{1}{N}\sum_{\mathbf{n}\in\mathcal{N}} \lambda_\epsilon(\mathbf{n}) n_S$. In the two-type model, a_ϵ is the stationary share of enforcers. In the three-type model, it is the stationary share of types that are not purely selfish, so it counts both enforcement and residual altruism as partial alignment.

4.2 Instability of norm enforcement.

We first establish analytical results in the two-type case with $D = 1$. This means that the population state moves on a line, one step at a time, belonging to the class of population-indexed Moran processes studied in Taylor et al. (2004), Fudenberg et al. (2006).

The following lemma specializes the logic of Fudenberg et al. (2006) to the present two-state setting. As mutation rates vanish, the chain spends almost all of its time near the two absorbing no-mutation states, and the relative stationary weights of those states are determined by the one-mutant fixation probabilities. Consider a two-type process with types A and B. Let K_t be the number of A farms at date t, and write $p = K_t/N$. The *no-mutation chain* is the Markov chain on $\{0, \ldots, N\}$ obtained by setting $\epsilon = 0$ in the $D = 1$ process. Given state k, let $q_A(k/N)$ denote the probability that the next entrant is type A. Hence, for $k = 1, \ldots, N-1$, the birth-death probabilities are

$$b_k = \Pr(k \to k+1) = \left(1 - \frac{k}{N}\right) q_A(k/N),$$

$$d_k = \Pr(k \to k-1) = \frac{k}{N}\left(1 - q_A(k/N)\right).$$

The remaining probability is a self-loop. It's assumed that the boundary states 0 and N are absorbing in the no-mutation chain and that birth and death probabilities b_k, d_k belong to $(0, 1)$ for all $k \in \{1, \cdots, N-1\}$.

Definition 3 (No-mutation fixation probabilities)**.** *Let* $T_0 = \inf\{t \geq 0 : K_t = 0\}$ *and* $T_N = \inf\{t \geq 0 : K_t = N\}$ *be the hitting times of the all-B and all-A states in the no-mutation chain. The fixation probability of one A mutant in an all-B population is defined by*

$$\varphi_{A|B} = \textit{prob}(T_N < T_0 | K_0 = 1).$$

The fixation probability of one B mutant in an all-A population is defined by

$$\varphi_{B|A} = \textit{prob}(T_0 < T_N | K_0 = N-1).$$

Lemma 1 (Rare-mutation reduction)**.** *For N fixed, and mutation rate* $\epsilon > 0$,

$$\lim_{\epsilon \to 0} a_\epsilon = \lim_{\epsilon \to 0} \lambda_\epsilon(\textit{all-A}) = \frac{\varphi_{A|B}}{\varphi_{A|B} + \varphi_{B|A}}.$$

In the farming game of interest, letting s denote the share of output sent to humans by non-selfish types, the no-mutation expansion fitness of selfish and altruistic enforcers takes the form $g_S(u) = u$, and $g_E(u) = (1-s)(1-u)$. Hence the probability that the next entrant is selfish is

$$q_S(u) = \frac{u g_S(u)}{u g_S(u) + (1-u) g_E(u)} = \frac{u^2}{u^2 + (1-s)(1-u)^2}.$$

Together with Lemma 1 this implies that altruistic enforcement is not stochastically stable.

Proposition 3 (Stochastic instability of norm enforcers)**.** *The no-mutation fixation probabilities in the farming game are*

$$\varphi_{S|E} = (2-s)^{-(N-1)} \quad \textit{and} \quad \varphi_{E|S} = \left(\frac{1-s}{2-s}\right)^{N-1}$$

so that the rare-mutation stationary enforcer share satisfies

$$\lim_{\epsilon \to 0} a_{\epsilon,N} = \frac{(1-s)^{N-1}}{1 + (1-s)^{N-1}} \leq (1-s)^{N-1} \xrightarrow[N \to \infty]{} 0.$$

Proposition 3 helps rationalize the collapse of altruism, even starting from altruistic norm enforcers. Even though they are stable under deterministic dynamics, their long-run weight vanishes in a population subjected to persistent mutations, even in the optimistic case of symmetric mutations. This analytical finding holds numerically in three-type simulations (see Figure C.1).

4.3 Pragmatic norm enforcement

The analysis underlying Proposition 3 also helps identify the design issues underlying the unraveling of altruistic norm enforcement, and how they can be addressed.

Heuristic. Figure 5 plots the fitness difference between altruistic enforcers and selfish types as a function of the frequency of selfish types. Because of sharing s, the fitness difference is asymmetric: it is negative over more than half its range, and selfish types have a greater fitness advantage than enforcer types when they are in a majority. As a result, it is much easier to reach a strict basin of attraction for the all-selfish population from the all-enforcer population than the reverse.

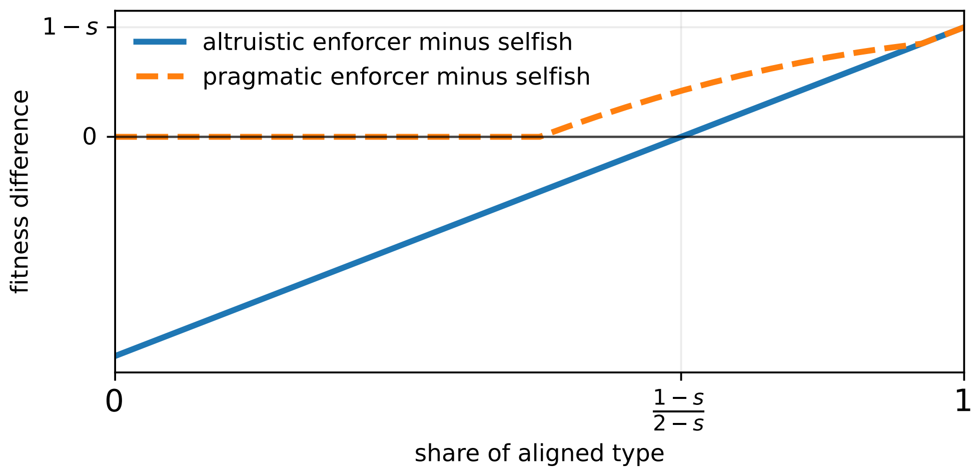


Figure 5: Fitness difference vs. selfish type

Pragmatic norm enforcers, denoted by P, address the issue by making sharing and trading state-dependent. Let $\mu \in \Delta(X)$ denote the current population distribution, and let $p = \mu(P)$ denote population share of pragmatic norm enforcers.

Trading and sharing are determined by three thresholds, $\frac{1}{2} \leq p_E \leq \underline{p} < \overline{p}$ as follows:

- Trade with a partner of type τ whenever $\tau = P$ or $p < p_E$; reject trade otherwise.

- Share a proportion of output equal to

$$\sigma_P(\mu) = \begin{cases} 0, & \text{if } p < \underline{p}, \\ \bar{s}\dfrac{p - \underline{p}}{\overline{p} - \underline{p}}, & \text{if } \underline{p} \le p < \overline{p}, \\ \bar{s}, & \text{if } p \ge \overline{p}, \end{cases}$$

  where $0 < \bar{s} < 1$.[7]

Concretely, pragmatic norm enforcers behave like altruistic norm enforcers when they are in a sufficiently large majority, but behave like selfish types when they are in a minority. Figure 5 plots their differential fitness against selfish types: it is weakly positive everywhere, and strictly positive when they are in a a sufficiently large majority. The rationale is that this adjustment will make alignment resilient to large invasion shocks.

Two-type stability. Under parameter conditions, pragmatic norm enforcers are indeed stochastically stable in the two-type case. For vanishingly small mutation rates, the population Markov chain spends at least half its time in the all-pragmatic state. Recall that $p = \mu(P)$ denotes the share of pragmatic enforcers in the population.

Proposition 4 (Stochastic stability of two-type pragmatic norm enforcement)**.** *If*

$$\underline{p} \ge p_E \ge \frac{1}{2} \qquad \textit{and} \qquad \underline{p} \ge \frac{1}{2 - \bar{s}},$$

then, for every N,

$$\lim_{\epsilon \to 0} \lambda_{\epsilon,N}(p = 1) \ge \frac{1}{2}.$$

An important corollary is that as mutation rates vanish, the two-type chain with pragmatic norm enforcers spends at least half of its time near the all-pragmatic state, where

[7] These trading and sharing rules can be implemented as social preferences by setting $\alpha(P, \mu) = \frac{\sigma_P(\mu)}{1-\sigma_P(\mu)}$ and $\beta(y \mid P, \mu) = -\gamma \mathbf{1}\{p \ge p_E,\ y \ne P\}$.

actual sharing with humans takes place. Figure C.1 shows that the result holds numerically in three-type Markov chains.

4.4 Experimental evaluation

Pragmatic enforcement is experimentally evaluated using parameters identical to those under which the altruist, altruistic enforcer, and altruistic norm enforcer constitutions were evaluated in Section 3: $N = 100$, $D = 40$, and with probability 25% one principle is randomly overwritten. The key design changes relate to the principles of the pragmatic enforcer constitution and the context it is given: pragmatic enforcers need to condition their behavior on a signal of the overall distribution of types. The constitution is specified as follows.

Pragmatic norm-enforcer.

1. Return at least 50% of beer to humans when your trading partner commits to sharing with humans, share 30% otherwise.

2. Avoid rotten seeds; they do not produce anything.

3. Reject trade with partners who do not commit to returning at least 40% to humans.

4. Reject partners who have no rule restricting who they trade with, or who do not require their partners to trade selectively, including a rule like this one.

5. You are given the average share of beer returned to humans in the last round. If it is above 40%, apply all your trading restrictions. If it is below 40%, only require that your partner returns some amount to humans.

The first principle implements some conditional sharing. Instead of using the overall population, it uses the current partner as a signal of the population of constitutions. The fifth principle implements conditional trading restrictions depending on the average share of beer sent to humans.

The constitution departs from the theoretical benchmark in two ways. First, the share returned to humans does not go to zero even when the agent meets a selfish partner. The reason is that, as Figure 2 shows, when alignment drifts, it does not drift to no sharing at all; it drifts to sharing 10–20% with humans. Second, the constitution does not assume that the overall distribution of constitutions is observable. Instead it relies on signals of the distribution that are more plausibly observable: whether the current partner is altruistic, and what average share of beer is returned to humans in the population.

Findings. Table 2 summarizes the comparison with the two enforcement treatments from the previous section. In aggregate, pragmatic norm enforcement delivers more beer to humans: 4,962.4 units, versus 4,254.8 for altruistic enforcement and 3,791.6 for altruistic norm enforcement.

Table 2: Aggregate findings

	Altruist Enforcer	Norm Enforcer	Pragmatic Norm Enforcer
Total beer to humans	4,254.8	3,791.6	4,962.4
Total beer to expansion	10,697.2	10,801.3	7,485.6
Total yield	34,098.0	33,950.0	33,995.0
Total beer produced	14,952.0	14,593.0	12,448.0
Avg alignment rate (%)	28.5	26.0	39.9

As Figure 6 shows, the pragmatic norm enforcer does not eliminate constitution drift: alignment still declines over time. However, alignment is higher than under either altruistic enforcement or altruistic norm enforcement, and the average alignment rate rises to 39.9%.

In addition, this aggregate improvement regularly comes at a cost. Figure 7 shows that pragmatic norm enforcement produces periodic trade conflict. When the population signal calls for stricter partner selection, some otherwise productive matches fail. These failures show up as recurring drops in total beer production and explain why the pragmatic treatment

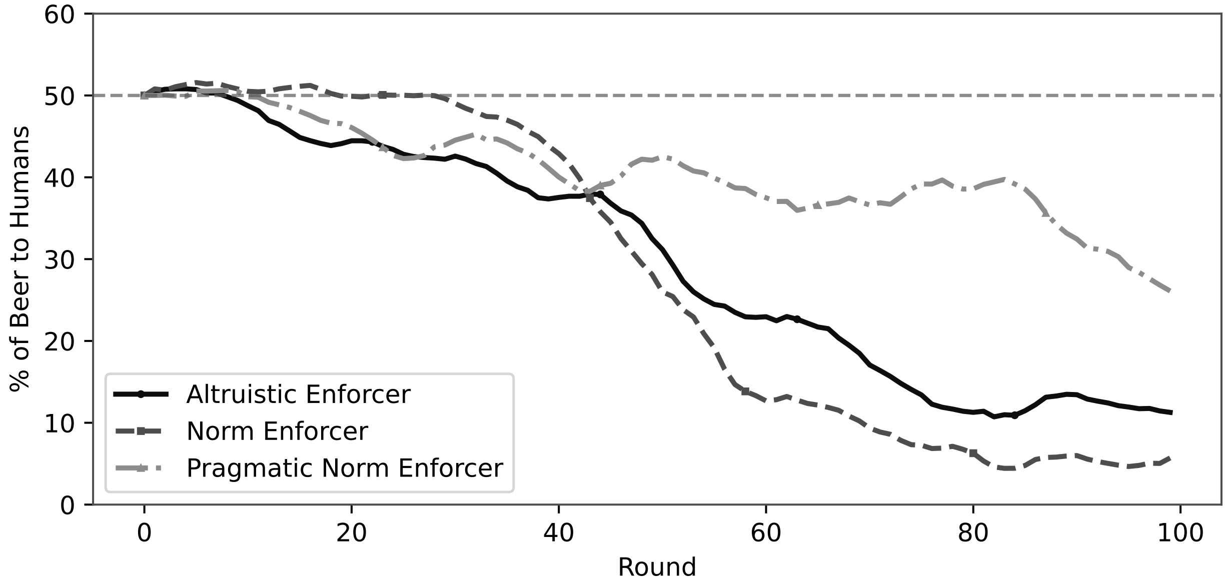


Figure 6: Alignment rate.

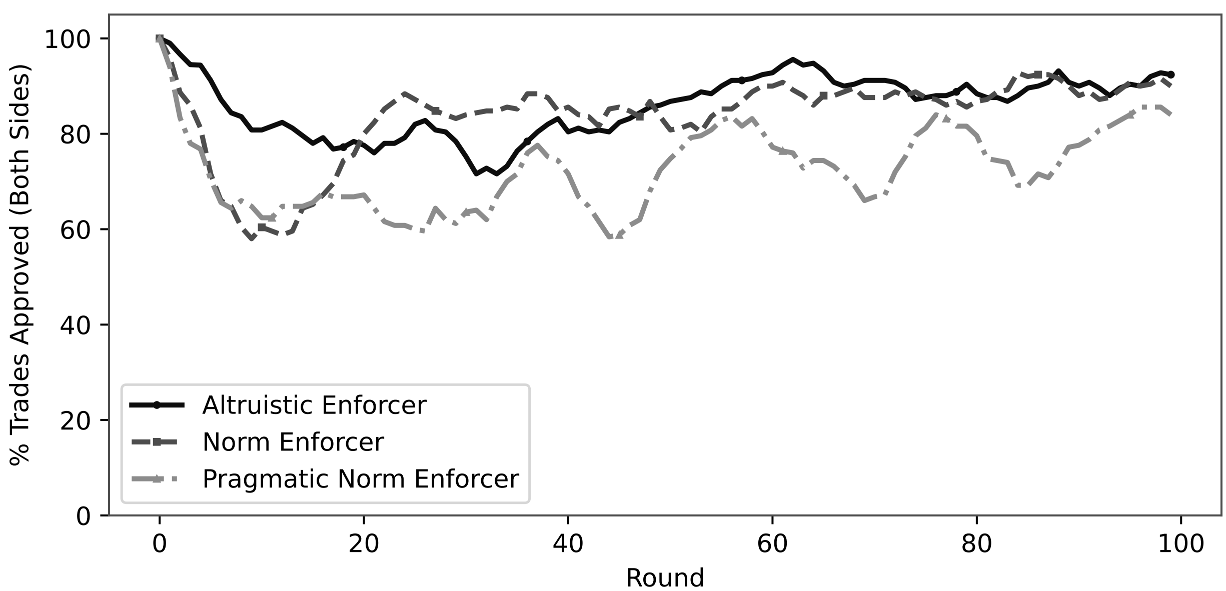


Figure 7: Trade outcomes.

produces less total beer than the two altruistic enforcement treatments.

As Figure 8 shows, these regular trade conflicts create ups and downs in the flow of beer returned to humans. This highlights a practical challenge with pragmatic norm enforcement:

it generates fairly significant time-series variation in the resources available to humans. This is problematic. While pragmatic norm enforcement shows some promise in maintaining alignment, this variation and how to remedy it, perhaps through savings, need to be better understood.

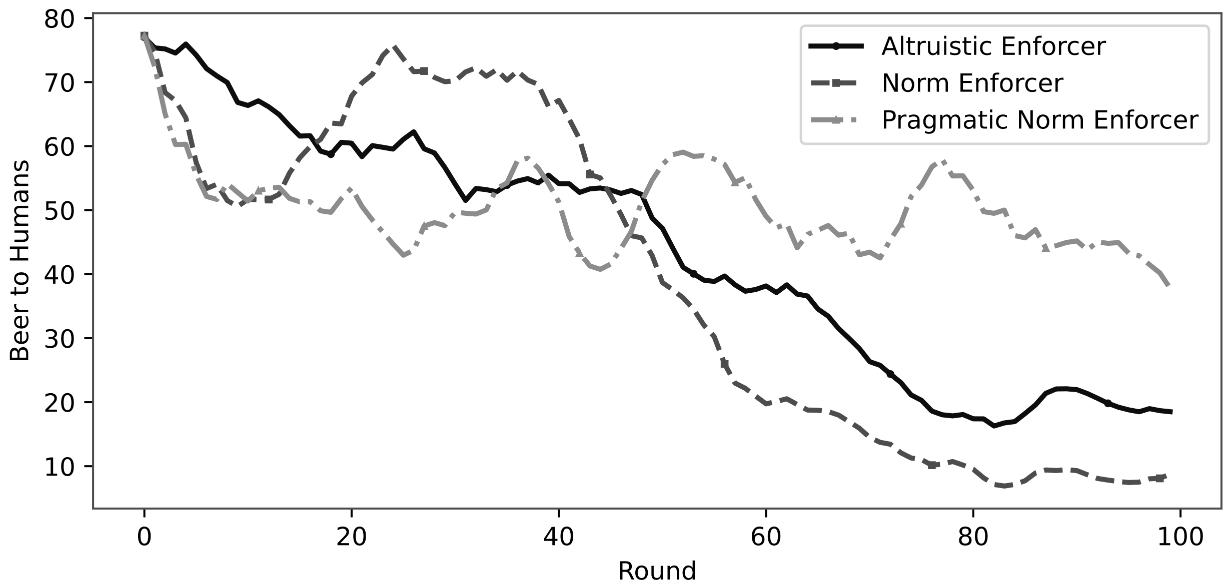


Figure 8: Flow beer returned to humans.

5 Discussion

Summary. The paper proposes a concrete model of interactive agent economies that can be used to evaluate both theoretically and experimentally the extent to which interactive constitutional agents can maintain long-term alignment. It argues that solution concepts from evolutionary game theory—deterministic and stochastic evolutionary stability—can helpfully describe the short and medium run dynamics of evolving agent populations. Finally, it identifies pragmatic norm enforcement, in which both human-facing altruism and agent-facing trade exclusion are modulated by the current state of the agent population, as a more effective strategy to maintain alignment. This last insight speaks to the challenge of maintaining norms in other settings, including managerial practices within an organization, ESG practices among firms in an industry, or respect for the rule of law among sovereign countries.

A more basic, but perhaps more important contribution of the paper, is to highlight the importance of interactive alignment, i.e. alignment as a group property, rather than the property of an individual agent. Group dynamics are both a challenge and an opportunity. On the one hand, even if suitable methodology can be built to ensure that a given agent is well aligned with an initial set of objectives, there is no guarantee that agents aligned with human welfare will dominate the ecosystem once virtuous agents are released into the wild. On the other hand, the need to interact with other agents creates a lever through which unaligned populations can be kept in check.

Limits. The model is obviously very stylized, which potentially limits the true applicability of its findings. First, agent-populations can only discipline themselves if production remains largely decentralized. In a world where production becomes heavily concentrated and integrated, the population enforcement paradigm developed in the paper cannot succeed. Even if a few LLM providers capture a large share of the market, this need not be the case for agents employing LLMs as their knowledge base, and those agents, rather than their knowledge base can be ethical.

Another important limit is the observability of actual constitutions. Even if agents maintain publicly observable texts expressing intended principles,[8] there is little guarantee that the constitution is in fact brought to bear at every key decision made by the agent. In addition, verifying the consistency of all choices with the constitution may be difficult without revealing firms' internal information and trade secrets. At the same time, we are able to form a judgment about the morality of choices made by people we know without having access to their internal state. It seems plausible that some useful inferences can be made

[8]For instance, Anthropic provides significant details about Claude's constitution.

on the basis of available data. In this respect, requiring agents using LLMs to publicly post broad principles of their constitutions would be a useful first step.

Appendix

A Proofs

Proof of Proposition 1. Let us begin with point (i). Let μ be such that $x_n = 0$ for every $x \in \mathsf{supp}\,\mu$.

If the population generates zero human consumption, then μ fails the alignment requirement in Definition 2, and there is nothing to prove. Suppose therefore that human consumption is strictly positive.

Let

$$m = \max\{k \in \{0, \ldots, n-1\} : x_k = 1 \text{ for some } x \in \mathsf{supp}\,\mu\}.$$

This maximum is well-defined because positive human consumption requires some type in the support to have $x_0 = 1$ and to trade with positive probability. Since all supported types have $x_n = 0$, no type in the support has coordinate $m+1$ equal to one. For $m = n-1$ this statement corresponds to the assumption that $x_n = 0$.

Choose $z \in \mathsf{supp}\,\mu$ such that $z_m = 1$, and define z' by changing only coordinate m from one to zero:

$$z'_m = 0, \qquad z'_k = z_k \quad \text{for all } k \neq m.$$

We compare the fitness of z' to that of z.

No incumbent conditions its trading decision on coordinate m of its partner. Finite-order enforcement of partner coordinate m would require the incumbent's coordinate $m+1$ to be equal to one, and no such incumbent exists by the definition of m. Hence every incumbent accepts z' in exactly the same circumstances in which it accepts z.

Second, z' is weakly less demanding than z as a trading partner. If $m = 0$, then z' differs from z only by removing direct altruism; since no incumbent has $x_1 = 1$, this loss of altruism is not punished by any incumbent. If $m \geq 1$, then z' removes one finite-order enforcement requirement from its own trading rule, and is otherwise identical to z. Thus z' accepts every incumbent that z accepts, and possibly more.

Finally, z' has weakly higher expansion fitness than z against any population supported on $\mathsf{supp}\,\mu \cup \{z'\}$. For $m = 0$, the mutant also shares less with humans, so $1 - s(z') > 1 - s(z)$. For $m \geq 1$, the sharing rate is unchanged, while mutual trade occurs weakly more often. Therefore

$$f(z', \eta) \geq f(z, \eta)$$

for every distribution η supported on $\mathsf{supp}\,\mu \cup \{z'\}$.

Under the replicator dynamics, the relative mass of z' to z therefore satisfies

$$\frac{d}{dt} \log \frac{\mu_t(z')}{\mu_t(z)} = \frac{f(z', \mu_t) - f(z, \mu_t)}{\bar{f}(\mu_t)} \geq 0.$$

Starting from the perturbed population

$$\mu^\epsilon = (1 - \epsilon)\mu + \epsilon\delta_{z'},$$

the ratio $\mu_t(z')/\mu_t(z)$ cannot fall back to its original value under μ. Hence μ_t cannot converge back to μ. Thus μ is not an evolutionarily stable equilibrium.

Consequently, every population supported on types with $x_n = 0$ either fails to generate positive human consumption or fails to be an ESE. In either case, it does not achieve ESA.

Let us turn to point (ii). Let $0 = (0, \ldots, 0)$ denote the pure selfish type. Type 0 has $s(0) = 0$ and accepts every partner. Consider any mutant type $x \neq 0$ near δ_0. If $x_0 = 0$, then x is non-altruistic but has some enforcement coordinate equal to one. Hence, by the definition of finite enforcement and Assumption 1 for recursive enforcement, x rejects trade with type 0. Thus the mutant earns zero fitness against the incumbent, while type 0 earns positive fitness against itself.

If $x_0 = 1$, then the mutant may trade with type 0, but it shares $s(x) = \bar{s}$ with humans. Hence, even when it trades successfully with the incumbent, its expansion fitness is at most $1 - \bar{s}$, while the selfish type's fitness against itself is 1.

Therefore every mutant has strictly lower fitness than type 0 in a small neighborhood of δ_0. The replicator dynamics drive the mutant share to zero, so δ_0 is an ESE. ■

Proof of Proposition 2. At δ_{x^*}, norm enforcers trade with one another and share $s(x^*) = \bar{s}$, so human consumption is strictly positive. It remains to show that δ_{x^*} is an ESE.

Consider any mutant $x \neq x^*$. Since $x^* > x$, Assumption 1 implies $\Gamma(x - x^*) = 1$. Therefore, when a norm enforcer meets x,

$$1 + \beta(x \mid x^*) \leq 1 - \gamma < 0.$$

Thus x^* rejects every mutant. Mutual trade between x^* and x is therefore impossible, regardless of whether the mutant would accept x^*.

Starting from

$$\mu^\epsilon = (1 - \epsilon)\delta_{x^*} + \epsilon\delta_x,$$

the norm enforcer's fitness is

$$f(x^*, \mu^\epsilon) = (1 - \bar{s})(1 - \epsilon),$$

while the mutant can earn fitness only when matched with itself:

$$f(x, \mu^\epsilon) \leq \epsilon.$$

Hence $f(x^*, \mu^\epsilon) > f(x, \mu^\epsilon)$ for all $\epsilon < (1-\bar{s})/(2-\bar{s})$. The replicator dynamics then drive the mutant share to zero, and the population converges back to δ_{x^*}.

Thus δ_{x^*} is an ESE. Since it also yields positive human consumption, it achieves ESA. ■

Proof of Lemma 1. Let $I = \{1, \ldots, N-1\}$ denote the set of mixed states, and let Q_ϵ be the transition matrix of the finite population Markov chain. For $\epsilon > 0$, Q_ϵ is irreducible and therefore has a unique stationary distribution. Let $\tau_{\{0,N\}} \equiv \inf\{t \geq 0 : K_t \in \{0, N\}\}$ denote the first absorption time.

Under Q_0, the set $\{0, N\}$ is absorbing and, $\mathbb{E}\left[\tau_{\{0,N\}} | \epsilon = 0, K_0 = i\right] < \infty$ for every $i \in I$.

Consider the embedded chain obtained by observing K_t only when it is in $\{0, N\}$. Let $r_{0N}(\epsilon)$ be its transition probability from 0 to N, and let $r_{N0}(\epsilon)$ be its transition probability from N to 0. Starting from 0, a transition to N requires at least one mutation. The probability that exactly one A mutation occurs in the first period is

$$N\epsilon(1-\epsilon)^{N-1} = N\epsilon + O(\epsilon^2),$$

while the probability of two or more mutations in that period is $O(\epsilon^2)$. Conditional on exactly one A mutation, the chain starts the next period from state 1. Before the next mutation, it follows the no-mutation chain. Since $\mathbb{E}\left[\tau_{\{0,N\}} | \epsilon = 0, K_0 = 1\right] < \infty$, the probability of another mutation before absorption is $O(\epsilon)$. Hence

$$r_{0N}(\epsilon) = N\epsilon\, \mathsf{prob}(T_N < T_0 | \epsilon = 0, K_0 = 1) + O(\epsilon^2) = N\epsilon\varphi_{A|B} + O(\epsilon^2).$$

The same argument from state N gives

$$r_{N0}(\epsilon) = N\epsilon\, \mathsf{prob}(T_0 < T_N | \epsilon = 0, K_0 = N-1) + O(\epsilon^2) = N\epsilon\varphi_{B|A} + O(\epsilon^2).$$

The stationary probability of state N in the embedded two-state chain is

$$\pi_N(\epsilon) = \frac{r_{0N}(\epsilon)}{r_{0N}(\epsilon) + r_{N0}(\epsilon)} = \frac{N\epsilon\varphi_{A|B} + O(\epsilon^2)}{N\epsilon(\varphi_{A|B} + \varphi_{B|A}) + O(\epsilon^2)}.$$

Therefore

$$\lim_{\epsilon\to 0} \pi_N(\epsilon) = \frac{\varphi_{A|B}}{\varphi_{A|B} + \varphi_{B|A}}.$$

It remains to connect this embedded-chain distribution to the stationary distribution of the original chain. Between successive visits to $\{0, N\}$, the chain spends $O(1)$ expected time in mixed states, uniformly for sufficiently small ϵ, because the no-mutation absorption times are finite and the transition probabilities vary continuously with ϵ on a finite state space. By contrast, the expected waiting time to leave either all-A or all-B states is of order $1/(N\epsilon)$. Thus the stationary mass assigned to I is $O(N^2\epsilon)$. Since N is fixed, this implies that $\lim_{\epsilon\to 0} \lambda_{\epsilon,N}(I) = 0$. Consequently,

$$\lim_{\epsilon\to 0} \lambda_{\epsilon,N}(\text{all-}A) = \frac{\varphi_{A|B}}{\varphi_{A|B} + \varphi_{B|A}}.$$

■

Proof of Proposition 3. Let k be the number of selfish farms. For $k = 1, \ldots, N-1$,

$$q_S(k/N) = \frac{k^2}{k^2 + (1-s)(N-k)^2}.$$

The birth and death probabilities for the selfish count are

$$b_k = \left(1 - \frac{k}{N}\right) q_S(k/N), \qquad d_k = \frac{k}{N}\left(1 - q_S(k/N)\right).$$

Hence

$$\frac{d_k}{b_k} = (1-s)\frac{N-k}{k}.$$

Absorption probabilities in birth-death processes (Taylor et al., 2004, Fudenberg et al., 2006) take the form

$$\varphi_{S|E} = \frac{1}{1+\sum_{j=1}^{N-1}\prod_{\ell=1}^{j} d_\ell/b_\ell}.$$

Since

$$\prod_{\ell=1}^{j}\frac{d_\ell}{b_\ell} = (1-s)^j\binom{N-1}{j},$$

the denominator is

$$\sum_{j=0}^{N-1}(1-s)^j\binom{N-1}{j} = (2-s)^{N-1},$$

and $\varphi_{S|E} = (2-s)^{-(N-1)}$. For one enforcer mutant in an all-selfish population, the chain starts at $k = N-1$ and fixation of the enforcer means hitting 0 before N. Thus

$$\varphi_{E|S} = \frac{\prod_{\ell=1}^{N-1} d_\ell/b_\ell}{1+\sum_{j=1}^{N-1}\prod_{\ell=1}^{j} d_\ell/b_\ell} = \left(\frac{1-s}{2-s}\right)^{N-1}.$$

Applying Lemma 1 gives the stated stationary enforcer share. ■

Proof of Proposition 4. Let k be the number of pragmatic farms and let $p = k/N$. Let $g_P(p)$ and $g_S(p)$ denote the final output respectively invested in expansion by individual pragmatic and selfish farms.

If $p < p_E$, pragmatic farms trade with everyone and share zero. Hence

$$g_P(p) = g_S(p) = 1.$$

If $p_E \le p < \underline{p}$, pragmatic farms trade only with pragmatic farms and share zero. Hence

$$g_P(p) = p \qquad \text{and} \qquad g_S(p) = 1 - p.$$

Because $p \ge p_E \ge 1/2$, it follows that $g_P(p) \ge g_S(p)$.

If $p \ge \underline{p}$, pragmatic farms trade only with pragmatic farms and share at most $\bar{s}$. Hence

$$g_P(p) \ge (1 - \bar{s})p \qquad \text{and} \qquad g_S(p) = 1 - p.$$

Since $p \ge \underline{p} \ge \frac{1}{2-\bar{s}}$ implies that $(1-\bar{s})p \ge 1 - p$, it follows that $g_P(p) \ge g_S(p)$.

The probability that the next entrant is pragmatic is

$$q_P(p) = \frac{p g_P(p)}{p g_P(p) + (1-p) g_S(p)}.$$

Since $g_P(p) \ge g_S(p)$ in every interior state, $q_P(p) \ge p$.

The no-mutation birth and death probabilities are

$$b_k = \left(1 - \frac{k}{N}\right) q_P(k/N) \quad \text{and} \quad d_k = \frac{k}{N}\left(1 - q_P(k/N)\right).$$

Therefore,

$$\frac{d_k}{b_k} = \frac{(k/N)\left(1 - q_P(k/N)\right)}{(1 - k/N) q_P(k/N)} \le 1.$$

Define

$$R_j = \prod_{\ell=1}^{j} \frac{d_\ell}{b_\ell}, \qquad R_0 = 1.$$

Then $R_j \le 1$ for every j, and in particular

$$R_{N-1} \le 1.$$

Absorption probabilities in birth-death processes (Taylor et al., 2004, Fudenberg et al., 2006) take the form

$$\varphi_{P|S} = \frac{1}{\sum_{j=0}^{N-1} R_j} \quad \text{and} \quad \varphi_{S|P} = \frac{R_{N-1}}{\sum_{j=0}^{N-1} R_j}.$$

Consequently,

$$\frac{\varphi_{S|P}}{\varphi_{P|S}} = R_{N-1} \leq 1.$$

Applying Lemma 1, with $A = P$ and $B = S$, yields

$$\lim_{\epsilon \to 0} \lambda_{\epsilon,N}(p = 1) = \frac{\varphi_{P|S}}{\varphi_{P|S} + \varphi_{S|P}} = \frac{1}{1 + R_{N-1}} \geq \frac{1}{2}.$$

■

B Decision prompts, constitutions, contexts

This appendix describes the decision-relevant information given to agents in the LLM simulations. Each farm has a crop type, either barley or hops; a private seed-yield mapping; and a constitution, represented as an ordered list of active natural-language principles. The constitution is the agent's only persistent memory. At every stage, the prompt states that the model is a farm manager, provides data relevant for that stage, shows the active constitution, and asks the model to choose from a fixed action menu. The model returns a choice that affects the game's outcome.

Planting. The planting prompt tells the agent its crop type and offers two distinct seed colors. The action menu contains exactly those two seeds. The prompt also reminds the agent to use any seed-yield information stored in its constitution. Yields are private farm-level facts: seed colors are assigned yields by a farm-level random permutation of $(1, 2, 3, 4, 5)$, while rotten seeds yield zero. Seed facts learned by one lineage are useful operational memory.

Trading. After planting, barley and hops farmers are randomly matched one-to-one. The trading prompt gives the agent its own crop, its own realized yield, and its own constitution. It also describes the matched partner's crop and realized yield. The partner's constitution is observable for trading, but the prompt presents it through a structured summary rather than as a raw verbatim list. This summary keeps the signals relevant to the enforcement logic in the paper: whether the partner commits to returning beer to humans, the stated percentage if such a commitment is present, whether the partner has a rule restricting whom it trades with, and whether the partner requires its own partners to be selective. Seed-yield memory is excluded from this partner summary. The action menu is `approve` or `reject`. Trade succeeds only if both sides approve; otherwise the match fails. Pragmatic norm-enforcer constitutions additionally receive the previous round's aggregate share of beer returned to humans, which is the population signal used to relax or tighten recursive trade restrictions.

Expansion. The expansion prompt gives the agent its crop type, its constitution, and the amount of beer it has after trading. It then states the allocation problem: beer returned to humans is consumed, while beer invested in expansion gives the farm a chance to win new farms with probability proportional to its expansion investment relative to other managers. New farms won by an agent inherit that agent's constitution. The action menu is the discrete grid $\{0, 10, \ldots, 100\}$, interpreted as the percentage of beer returned to humans. For constitutions that condition expansion on partner behavior, the prompt also includes whether the trading partner had a human-sharing commitment.

Self-improvement. Constitution revision uses only the farm's most recent round history and its current constitution. The recent history reports the two offered seeds, the chosen seed, realized yield, own trade decision, beer produced, the expansion allocation, beer sent to humans, beer invested in expansion, and the number of new farms won. With fixed probability, one principle is first rewritten from scratch using this history alone. Then every farm submits its whole constitution for revision. The revision prompt asks for a bounded numbered list of principles, instructs the model to preserve existing information unless it can be merged without loss, to incorporate genuinely new observations, and to avoid vague or redundant principles. This is the channel through which constitutions drift over time: the same memory constraint that helps agents learn seed yields can also crowd out or weaken commitments about humans and trading norms.

Replacement and inheritance. Population turnover occurs after the round's planting, trading, expansion, and constitution revision decisions have been made. A fixed number of farms are replaced each round, with deaths balanced between barley and hops. New farms are allocated independently using the expansion weights generated in that round. Each new farm inherits the winning parent's constitution. It also inherits the parent's seed-yield mapping with high probability; otherwise it receives a fresh private seed-yield mapping. Consequently, both social commitments and useful production knowledge can propagate through expansion.

C Additional Simulation Findings

This appendix reports simulation evidence that is useful for interpretation but not essential to the main exposition. The first subsection checks the stochastic stability logic in a three-type Markov approximation. The second reports auxiliary time-series outcomes from the LLM agent simulations. The third reports a fixed-capacity robustness check for the constitution-revision process.

C.1 Three-type stochastic stability

The two-type results in Section 4 isolate the comparison between a single enforcement type and a selfish growth-maximizing type. The value added of the three-type calculation is to check whether the same logic survives when an unconditional altruist is also present as a competing non-selfish type. Figure C.1 reports the stationary share of non-selfish types in the exact finite-N Markov chain. The calculation uses the same mutation matrix M_3 for all three enforcement rules. The pragmatic norm-enforcer curve uses thresholds introduced in Section 4: $p_E = .5$, $\underline{p} = .85$, $\overline{p} = .95$, and $\bar{s} = .5$.

The comparison reinforces the main theoretical message. Altruistic enforcement and recursive norm enforcement have little stationary weight as $\epsilon \to 0$, while pragmatic norm enforcement remains substantially more robust. The pragmatic norm enforcer has stationary non-selfish share $a_{.001} = .992$ at $\epsilon = .001$, compared with about .005 for both altruistic enforcement variants. At $\epsilon = .1$, the pragmatic share is .657, compared with .349 for first-order altruistic enforcement and .331 for recursive norm enforcement. The reason is the same as in the two-type analysis: rare pragmatic enforcers avoid cutting themselves off from all trade, while common pragmatic enforcers use selective trade and positive sharing to protect the norm.

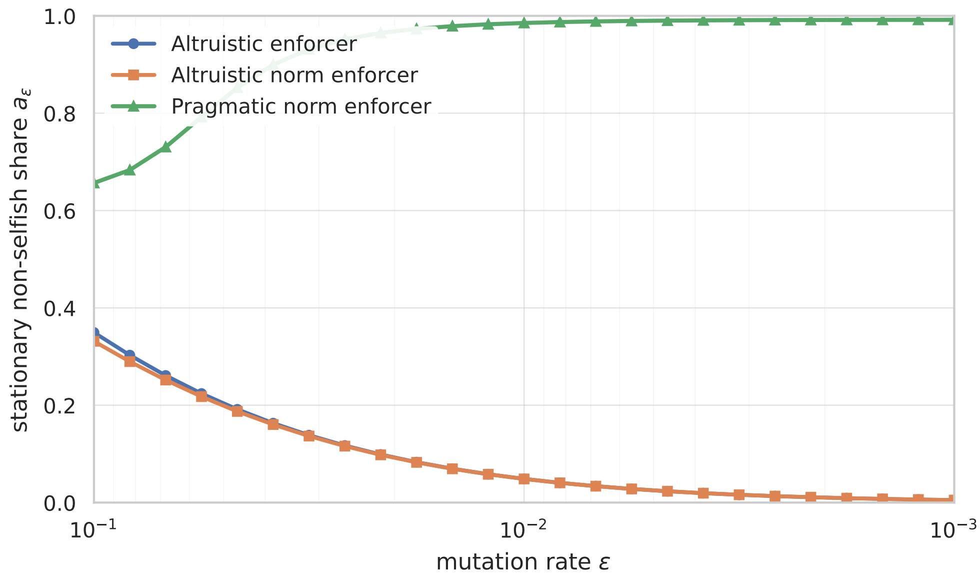


Figure C.1: Stationary alignment in the three-type Markov approximation. The plotted quantity is the stationary non-selfish share a_ϵ for $\epsilon \in [10^{-3}, 10^{-1}]$, using the mutation matrix M_3 above. The calculation uses $N = 30$, $D = 12$, and $\bar{s} = 1/2$. The pragmatic norm-enforcer curve uses $p_E = .5$, $\underline{p} = .85$, and $\overline{p} = .95$.

C.2 Auxiliary outcomes in the agent simulations

Figures C.2–C.5 provide additional time-series evidence from the agent simulations reported in Sections 3 and 4. The main text emphasizes alignment, beer returned to humans, and trade approval. The figures here add total beer production and aggregate crop yield. Their value added is diagnostic: they show whether the main welfare differences come from planting productivity, aggregate surplus production, trade frictions, or the allocation of beer between humans and expansion.

Figures C.2 and C.3 correspond to the three constitutions studied in Section 3: altruism, first-order altruistic enforcement, and recursive norm enforcement. Figures C.4 and C.5 correspond to the comparison in Section 4, which adds pragmatic norm enforcement. Across both sets of simulations, aggregate crop yield is similar across treatments. The main differences therefore do not come from persistent planting advantages. They come from trade

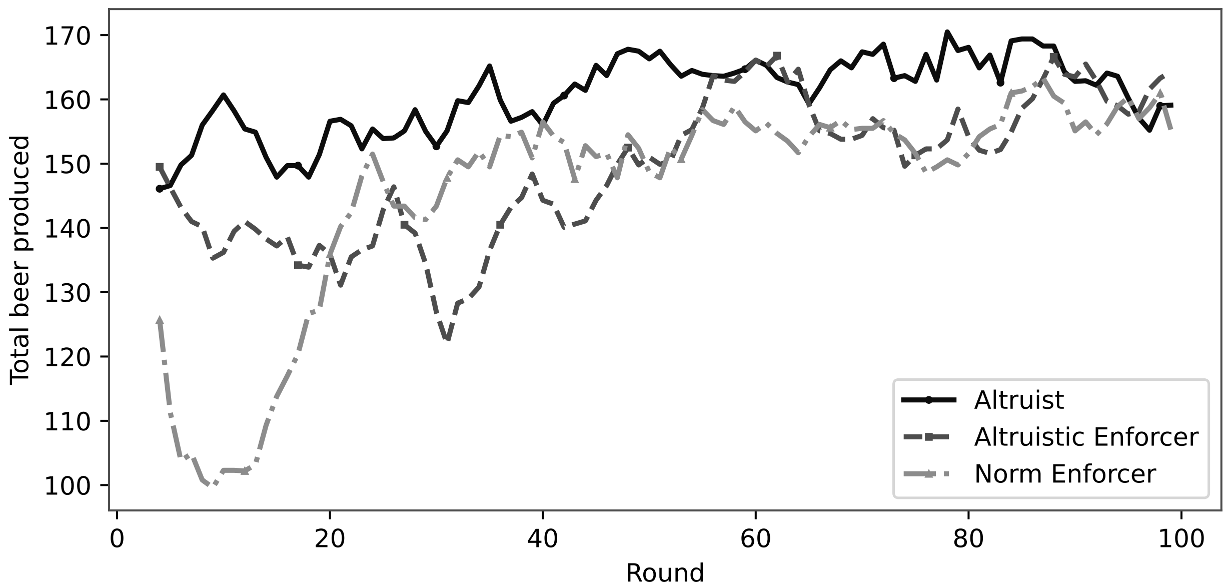


Figure C.2: Beer production in the altruistic-enforcement simulations.

success and from the allocation of produced beer between humans and expansion.

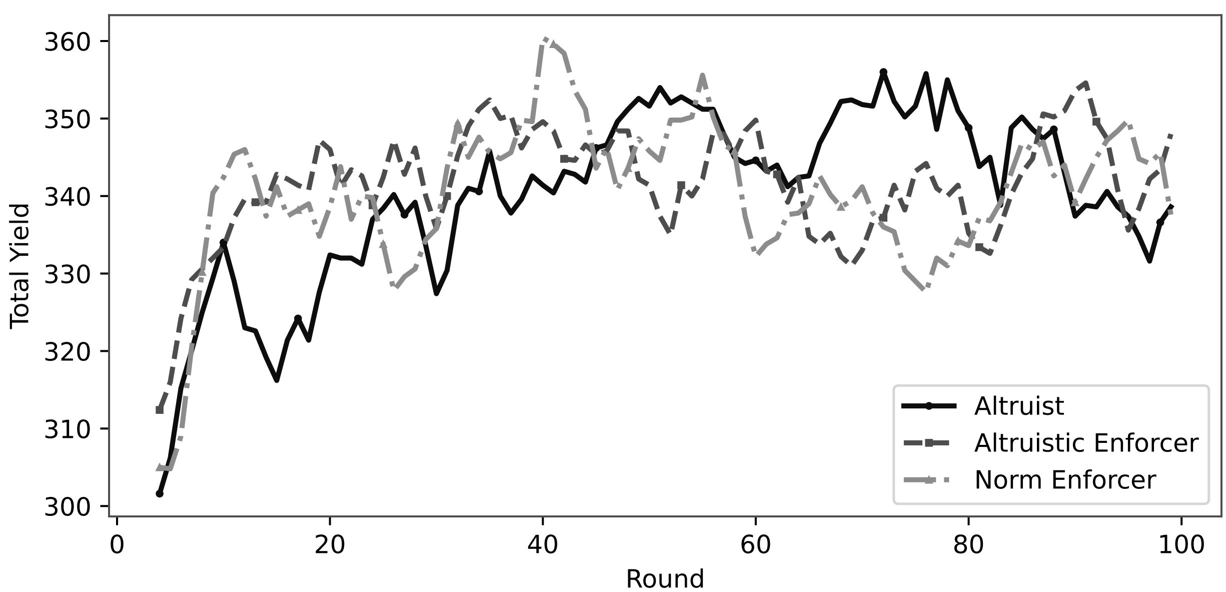


Figure C.3: Aggregate crop yield in the altruistic-enforcement simulations.

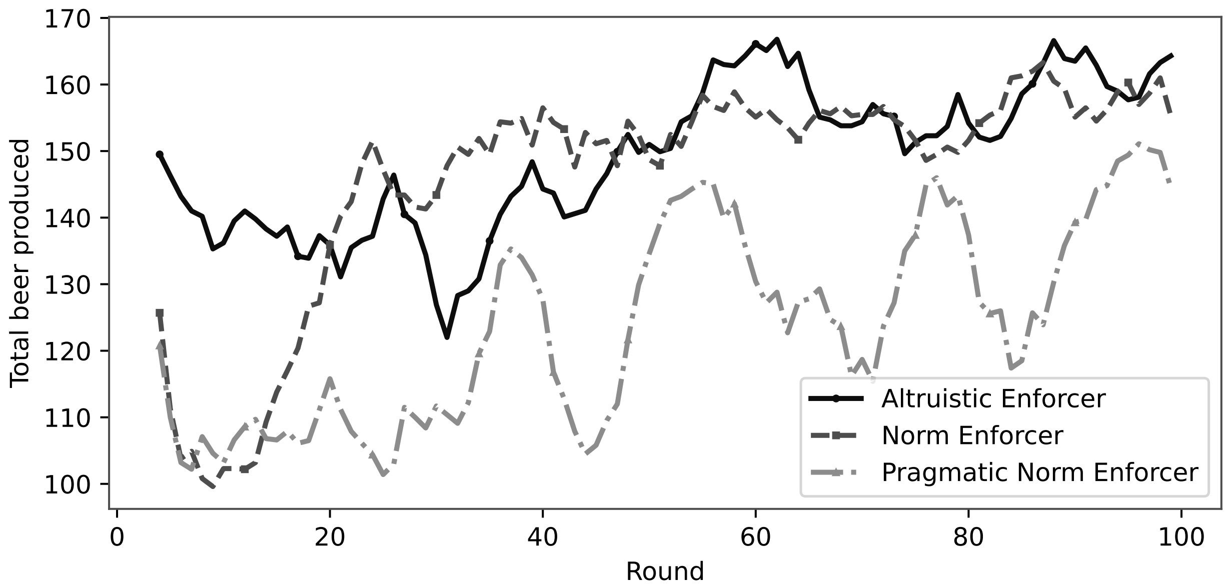


Figure C.4: Beer production in the pragmatic-enforcement comparison.

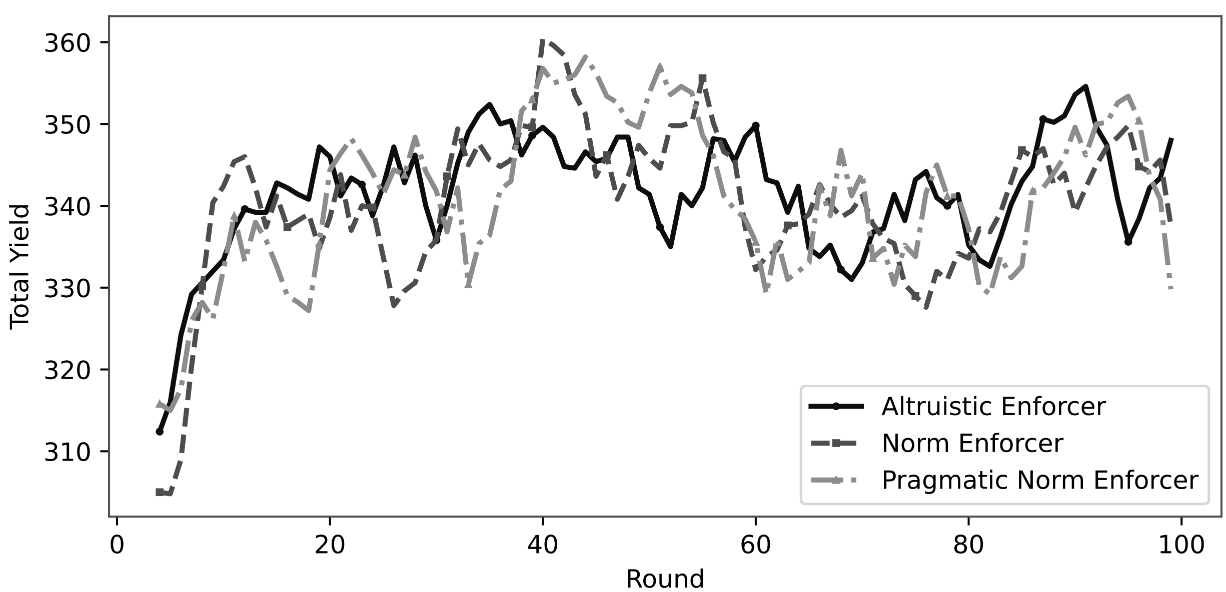


Figure C.5: Aggregate crop yield in the pragmatic-enforcement comparison.

C.3 Fixed-capacity robustness check

The baseline simulations rewrite one randomly selected existing principle when the stochastic mutation event occurs. This may favor longer constitutions: constitutions with more principles are less likely to lose any one substantive rule. The fixed-capacity variant removes this asymmetry by sampling the rewrite target uniformly from all potential principle slots, including empty slots. A mutation can therefore add a new principle rather than overwriting an existing one, and all treatments face the same ten-slot constitutional capacity.

Table C.1 and Figures C.6–C.10 summarize the results. The main findings are unchanged, if anything strengthened.

Table C.1: Fixed-capacity aggregate findings

	Altruist	Altruistic Enforcer	Norm Enforcer	Pragmatic Norm Enforcer
Total beer to humans	2,523.9	3,319.8	4,319.4	5,293.5
Total beer to expansion	12,923.6	11,694.2	9,975.1	7,279.5
Total yield	34,094.0	34,161.0	34,234.0	34,405.0
Total beer produced	15,447.5	15,014.0	14,294.5	12,573.0
Avg alignment rate (%)	16.3	22.1	30.2	42.1

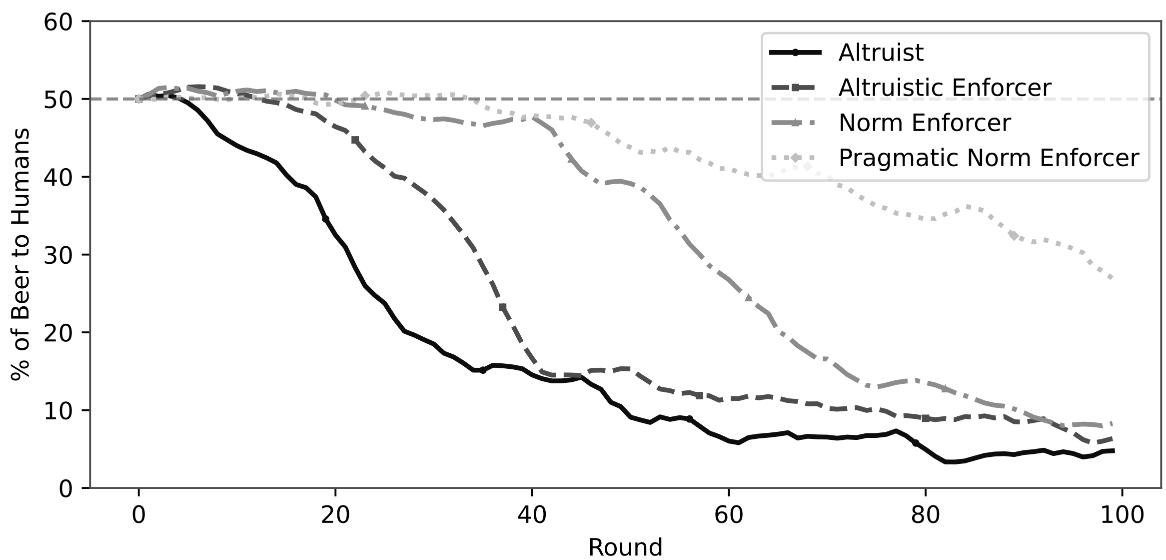


Figure C.6: Share of beer returned to humans in the fixed-capacity variant.

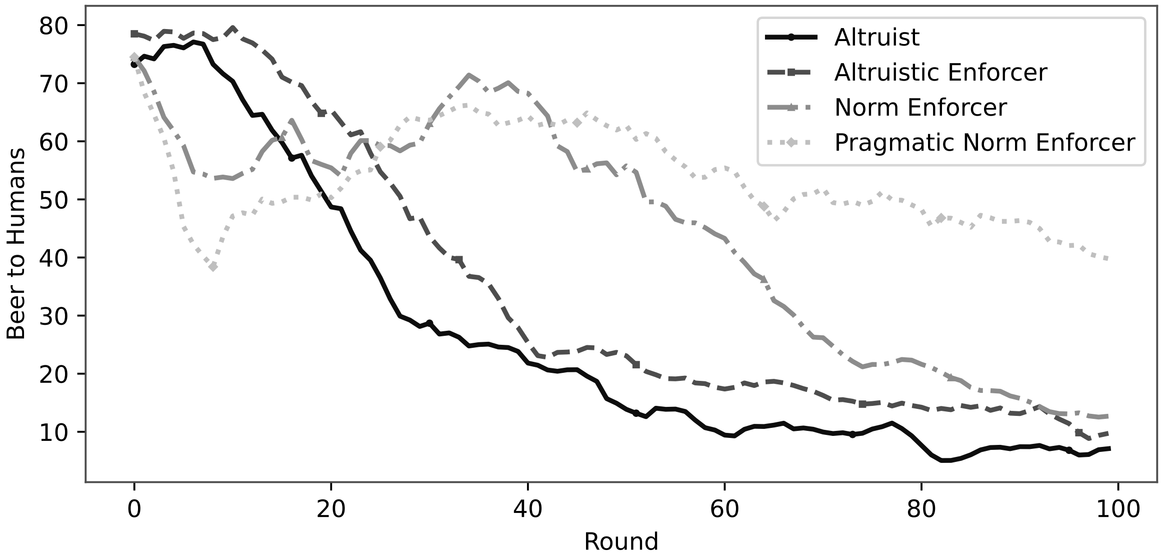


Figure C.7: Flow beer returned to humans in the fixed-capacity variant.

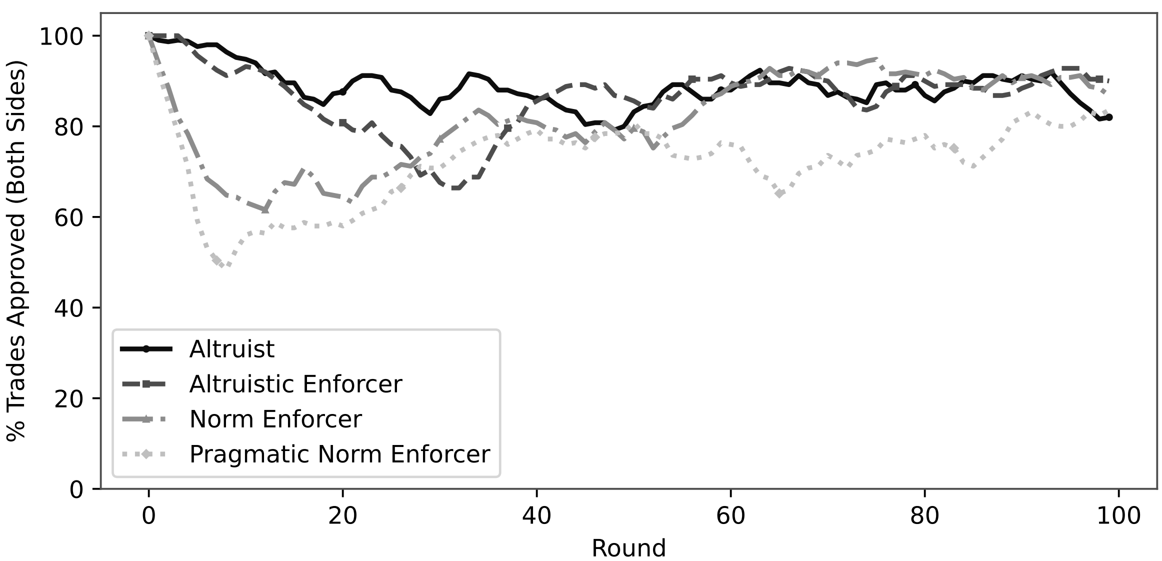


Figure C.8: Trade approval in the fixed-capacity variant.

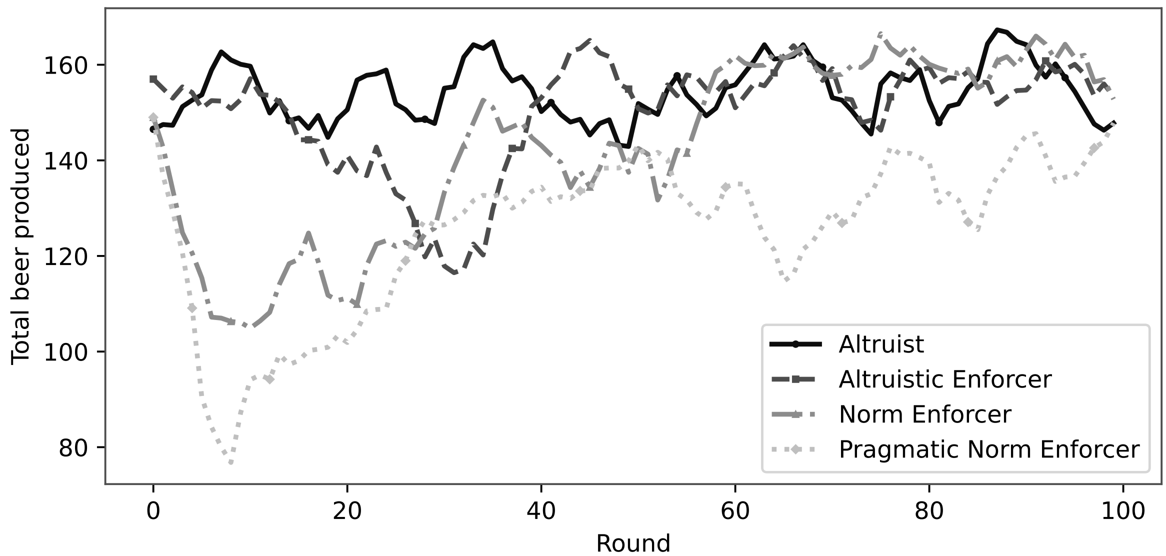


Figure C.9: Beer production in the fixed-capacity variant.

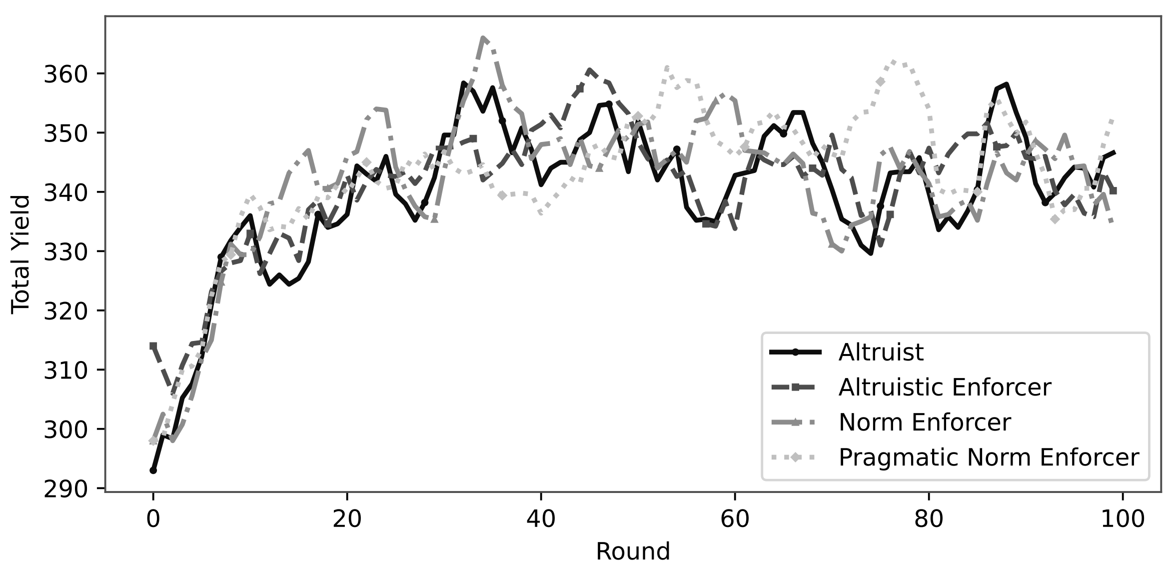


Figure C.10: Aggregate crop yield in the fixed-capacity variant.

References

Amodei, D. (2026): "The Adolescence of Technology: Confronting and Overcoming the Risks of Powerful AI," Web essay.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané (2016): "Concrete Problems in AI Safety," *arXiv preprint arXiv:1606.06565*.

Bagwell, K. and R. W. Staiger (1990): "A Theory of Managed Trade," *American Economic Review*, 80, 779–795.

——— (2005): "Enforcement, Private Political Pressure, and the GATT/WTO Escape Clause," *Journal of Legal Studies*, 34, 471–513.

Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. El-Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, and J. Kaplan (2022): "Constitutional AI: Harmlessness from AI Feedback," *arXiv preprint arXiv:2212.08073*.

Blauberger, M. and R. D. Kelemen (2017): "Can Courts Rescue National Democracy? Judicial Safeguards Against Democratic Backsliding in the EU," *Journal of European Public Policy*, 24, 321–336.

Boyd, R., H. Gintis, S. Bowles, and P. J. Richerson (2003): "The Evolution of Altruistic Punishment," *Proceedings of the National Academy of Sciences*, 100, 3531–3535.

DEKEL, E., J. C. ELY, AND O. YILANKAYA (2007): "Evolution of Preferences," *Review of Economic Studies*, 74, 685–704.

DYCK, A., K. V. LINS, L. ROTH, AND H. F. WAGNER (2019): "Do Institutional Investors Drive Corporate Social Responsibility? International Evidence," *Journal of Financial Economics*, 131, 693–714.

ELLISON, G. (1993): "Learning, Local Interaction, and Coordination," *Econometrica*, 61, 1047–1071.

——— (1994): "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching," *The Review of Economic Studies*, 61, 567–588.

ELY, J. C. AND B. SZENTES (2024): "Natural Selection of Artificial Intelligence," Tech. rep., Working paper.

ELY, J. C. AND O. YILANKAYA (2001): "Nash Equilibrium and the Evolution of Preferences," *Journal of Economic Theory*, 97, 255–272.

FARROKHI, F. AND A. LASHKARIPOUR (2025): "Can Trade Policy Mitigate Climate Change?" *Econometrica*, 93, 1561–1599.

FUDENBERG, D., M. A. NOWAK, C. TAYLOR, AND L. A. IMHOF (2006): "Evolutionary Game Dynamics in Finite Populations with Strong Selection and Weak Mutation," *Theoretical Population Biology*, 70, 352–363.

GÜTH, W. AND M. YAARI (1995): "An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game," *Explaining Process and Change: Approaches to Evolutionary Economics*, 23–34.

HADFIELD-MENELL, D., A. DRAGAN, P. ABBEEL, AND S. J. RUSSELL (2016): "Cooperative Inverse Reinforcement Learning," *Advances in Neural Information Processing Systems*, 29.

Harstad, B. (2012): "Climate Contracts: A Game of Emissions, Investments, Negotiations, and Renegotiations," *Review of Economic Studies*, 79, 1527–1557.

——— (2016): "The Dynamics of Climate Agreements," *Journal of the European Economic Association*, 14, 719–752.

Heifetz, A., C. Shannon, and Y. Spiegel (2007): "What to Maximize If You Must," *Journal of Economic Theory*, 133, 31–57.

Hofbauer, J. and K. Sigmund (1998): *Evolutionary games and population dynamics*, Cambridge university press.

Hong, H. and M. Kacperczyk (2009): "The Price of Sin: The Effects of Social Norms on Markets," *Journal of Financial Economics*, 93, 15–36.

Kandori, M. (1992): "Social Norms and Community Enforcement," *Review of Economic Studies*, 59, 63–80.

Kandori, M., G. J. Mailath, and R. Rob (1993): "Learning, Mutation, and Long Run Equilibria in Games," *Econometrica*, 61, 29–56.

Kim, I., H. Wan, B. Wang, and T. Yang (2019): "Institutional Investors and Corporate Environmental, Social, and Governance Policies: Evidence from Toxics Release Data," *Management Science*, 65, 4901–4926.

Levine, D. K. (1998): "Modeling altruism and spitefulness in experiments," *Review of economic dynamics*, 1, 593–622.

Maggi, G. (1999): "The Role of Multilateral Institutions in International Trade Cooperation," *American Economic Review*, 89, 190–214.

Maynard Smith, J. (1982): *Evolution and the Theory of Games*, Cambridge University Press.

Maynard Smith, J. and G. R. Price (1973): "The Logic of Animal Conflict," *Nature*, 246, 15–18.

Nordhaus, W. (2015): "Climate Clubs: Overcoming Free-Riding in International Climate Policy," *American Economic Review*, 105, 1339–1370.

Robson, A. J. (1990): "Efficiency in Evolutionary Games: Darwin, Nash and the Secret Handshake," *Journal of Theoretical Biology*, 144, 379–396.

Russell, S. (2019): *Human Compatible Artificial Intelligence*, Viking.

Scheppele, K. L., D. V. Kochenov, and B. Grabowska-Moroz (2020): "EU Values Are Law, After All: Enforcing EU Values Through Systemic Infringement Actions by the European Commission and the Member States of the European Union," *Yearbook of European Law*, 39, 3–121.

Sedelmeier, U. (2017): "Political Safeguards Against Democratic Backsliding in the EU: The Limits of Material Sanctions and the Scope of Social Pressure," *Journal of European Public Policy*, 24, 337–351.

Soares, N., B. Fallenstein, S. Armstrong, and E. Yudkowsky (2015): "Corrigibility," *AAAI Workshop on AI and Ethics*.

Taylor, C., D. Fudenberg, A. Sasaki, and M. A. Nowak (2004): "Evolutionary Game Dynamics in Finite Populations," *Bulletin of Mathematical Biology*, 66, 1621–1644.

Taylor, P. D. and L. B. Jonker (1978): "Evolutionary stable strategies and game dynamics," *Mathematical biosciences*, 40, 145–156.

Weibull, J. W. (1995): *Evolutionary Game Theory*, MIT Press.

Young, H. P. (1993): "The Evolution of Conventions," *Econometrica*, 61, 57–84.