An LLM-based Agent Simulation Approach to Study Moral Evolution

Zhou Ziheng[1*✉], **Huacong Tang**[1*], **Mingjie Bi**[2*], **Yipeng Kang**[2], **Wanying He**[2], **Fang Sun**[1], **Yizhou Sun**[1], **Ying Nian Wu**[1], **Demetri Terzopoulos**[1], **Fangwei Zhong**[3,2✉]

[1]University of California, Los Angeles [2]Beijing Institute for General Artificial Intelligence
[3]Beijing Normal University
[*]Equal contribution. [✉]Corresponding authors: josephziheng@ucla.edu, fangweizhong@bnu.edu.cn

Project Page Code

Abstract

The evolution of morality presents a puzzle: natural selection should favor self-interest, yet humans developed moral systems promoting altruism. Traditional approaches must abstract away cognitive processes, leaving open how cognitive factors shape moral evolution. We introduce an LLM-based agent simulation framework that brings cognitive realism to this question: agents with varying moral dispositions perceive, remember, reason, and decide in a simulated prehistoric hunter-gatherer society. This enables us to manipulate factors that traditional models cannot represent—such as moral type observability and communication bandwidth—and to discover emergent cognitive mechanisms from agent interactions. Across 20 runs spanning four settings, we find that cooperation and mutual help are the central drivers of survival, with universal and reciprocal morality exhibiting the most stable evolutionary outcomes while selfishness is strongly disfavoured. We also observe two central mechanisms that emerge from cognition, including a *cost of moral judgment* and a *self-purging effect* among selfish agents. We validate robustness across multiple LLM backbones, architecture ablations, and prompt sensitivity analyses. This work establishes LLM-based simulation as a powerful new paradigm to complement traditional research in evolutionary biology and anthropology, opening new avenues for investigating the complexities of moral and social evolution.

1 Introduction

The emergence and evolution of morality represents one of the most enduring puzzles in evolutionary biology and social sciences (Haidt, 2007; Greene, 2013). From an evolutionary standpoint, natural selection should favor individuals who maximize their reproductive success, often through selfish behaviors that increase resource acquisition at others' expense (Dawkins, 1976). Yet, humans and some other species have evolved complex moral systems that frequently promote cooperation, altruism, and other prosocial behaviors that can seemingly contradict individual fitness maximization (Tomasello, 2016). This apparent contradiction presents a profound scientific question: Under what conditions does morality provide an evolutionary advantage?

Prior research has approached this question through evolutionary game theory (Nowak, 2006; Axelrod and Hamilton, 1981), anthropological fieldwork on moral universals and cultural variation (Henrich, 2015; Curry et al., 2019a), and biological mechanisms including kin selection (Hamilton, 1964), reciprocal altruism (Trivers, 1971), and group selection (Wilson and Wilson, 2007). Moral frameworks have further identified patterns such as the expanding circle of moral concern (Singer, 1981) and fundamental moral dimensions (Haidt, 2007). While these approaches have yielded valuable insights, they share a fundamental limitation: they must abstract away the cognitive processes underlying moral behaviour, encoding agent strategies as fixed rules or payoff matrices. This prevents investigation of how cognitive factors—such as the ability to identify others' moral dispositions, memory of past interactions, and communication bandwidth—interact with environmental pressures to shape moral evolution.

Recent advances in LLMs present a novel methodological opportunity to address these limitations. LLM-based agent simulations can model entities with sophisticated cognitive architectures—including values, memory, perception, reasoning, and social dynamics—that generate emergent, complex behaviors (Park et al., 2023; Horton, 2023). This simulation paradigm allows us to observe interactions between moral cognition, behavior, and evolutionary outcomes under controlled conditions while providing rich, realistic psychological details that surpass traditional agent-based models (Aher

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28993–29046
July 2-7, 2026

et al., 2023). Nevertheless, the existing methods barely support research on morality evolution.

To address the gaps, our research advances the study of human evolution through three primary contributions: 1) Methodological: We pioneer a computational approach using LLM-based agents to bring psychological realism to evolutionary simulations, overcoming the limitations of traditional abstract models; 2) Programmatic: We release MoRE, an extensible agent framework, and Social-Evol, a versatile environment platform. Together, they enable multifaceted inquiry into social evolution—from norm formation to inter-group conflict—through both long-term and scenario-based simulation; 3) Empirical: We conduct experiments across multiple settings showing that cooperation is consistently favoured by selection, with universal and reciprocal morality exhibiting the most stable survival. We discover emergent mechanisms—including a self-purging effect among selfish agents and the decisive role of the cost of moral judgment—that are difficult to study with traditional approaches. We validate robustness across multiple LLM backbones, architecture ablations, and prompt sensitivity analyses.

2 Related Work

Evolutionary Origins of Morality Evolutionary biologists have proposed various mechanisms to explain how altruistic traits might evolve. Theories of kin selection (Hamilton, 1964) and reciprocal altruism (Trivers, 1971) show how limited forms of cooperation could evolve among relatives and repeated interaction partners. Recently, cultural group selection theories (Boyd et al., 2011; Henrich, 2015) explain how groups with stronger moral norms outcompeted others, leading to the genetic evolution of psychological predispositions supporting moral behavior. Evolutionary game theory provided mathematical frameworks demonstrating how cooperation can evolve under some strategies similar like previously mentioned mechanisms by showing strategies can yield higher payoffs than pure selfishness under specific conditions. Nowak (2006)'s "five rules for the evolution of cooperation" identifies key mechanisms: kin selection, direct reciprocity, indirect reciprocity, network reciprocity, and group selection.

Though these works provide great insights and a mathematical foundation for cooperation and moral evolution, they highly abstract away the rich complexity of human cognition and cooperation using mathematical models, blocking a full view of moral evolution dynamics.

Moral Frameworks Moral Foundations Theory identifies five moral dimensions: care/harm, loyalty/betrayal, authority/subversion, and sanctity/degradation (Haidt and Joseph, 2007). The Theory of Dyadic Morality (Gray et al., 2012) emphasizes harm as the root of morality, while Morality-as-Cooperation theory (Curry et al., 2019b) identifies seven cooperative behaviors as essential: helping kin, helping group members, reciprocating, being brave, deferring to superiors, dividing resources fairly, and respecting others' property. Other theories ground morality in distinctions between rules and conventions (Turiel, 1983), social-relational models (Fiske, 1992; Rai and Fiske, 2011), specific moral emotions (Rozin et al., 1999), or an ethic of care (Gilligan, 1982).

While existing theories offer rich descriptions of morality's central traits, our initial investigation desires a framework that organizes these traits into scalable "levels" for a more systematic analysis. The Expanding Circle Theory (Singer, 1981) provides an ideal structure for this purpose. Its concept of a "circle of concern" that expands from the self, to kin, and finally to society offers a clear, tiered structure for defining an agent's moral level. Furthermore, this progression provides a natural way to integrate key concepts of other theories: care and loyalty are paramount within the kin circle, while fairness and reciprocity are essential for the stability of the broader group circle. This concentric model has cross-cultural relevance due to its deep resonance with philosophical traditions like Confucianism (Fei, 1992), which also defines morality through the proper ordering of relational duties. Given its implementable structure and integrative power, we adopt the Expanding Circle as the primary theoretical foundation for our simulation.

LLM-Based Agent Simulation Recent advances in LLMs provide the methodological foundation for our approach to studying morality's evolution. LLM-based agent simulations can model entities with sophisticated cognitive architectures—including values, memory, perception, reasoning, and social dynamics—that generate emergent, complex behaviors (Park et al., 2023; Horton, 2023). LLM agents allow for controlled experimentation with variables that would be impossible to manipulate in real-world settings; they provide

Table 1: Comparison of Existing LLM-based Simulation Frameworks for Social Science.

Environment	Morality	Memory	Team Formation	Competition	Evolution	Engine
Park et al. (2023)	✗	Short/Long-term	✗	✗	✗	LLM
Li et al. (2023)	✗	Undefined	Multi-round Negotiation	✗	✗	LLM
Horton (2023)	✗	Undefined	✗	Explicit	✗	LLM
Aher et al. (2023)	✗	Short-term	✗	✗	✗	LLM
Wang et al. (2023)	✗	Short-term	✗	✗	✗	LLM
Huang et al. (2024)	✗	Short-term	Multi-round Negotiation	Implicit	✓	LLM/RL
Wang et al. (2024)	✗	Short-term	✗	✗	✗	LLM
Piao et al. (2025)	✗	Short/Long-term	Task-dependent	Task-dependent	✓	LLM
Guan et al. (2024)	✗	Short/Long-term	Multi-round Negotiation	Explicit	✓	LLM
Dai et al. (2024)	✗	Short/Long-term	Negotiation-based	Explicit	✓	LLM
Ours	✓	Short/Long-term	Multi-round Negotiation	Explicit	✓	LLM

transparent access to agents' decision-making processes; and they can simulate long timescales of social development (Piao et al., 2025).

Prior work has demonstrated the versatility of this approach. Park et al. (2023) created an interactive simulated small town where generative agents exhibited complex social behaviors, including relationship formation and collective problem-solving. Horton (2023) applied LLM agents to economic simulations, finding that agents replicate known economic phenomena while providing unprecedented access to reasoning processes. Aher et al. (2023) validated that LLM agent simulations can reproduce results from human behavioral experiments, showing its potential as a complementary methodology in social science research. However, existing environments lack the complete settings of agent morality, cooperative and competitive interactions, and evolution features, as demonstrated in Table 1. One relevant recent work is "Artificial Leviathan", which explores social order in LLM agent societies, while they assume all agents are inherently selfish and focus on how social order emerges from this assumption (Dai et al., 2024).

Therefore, by explicitly modeling agents with varying moral dispositions to better reflect the diversity of human moral psychology, our work represents the first systematic application of LLM-based simulations to investigate the evolution of morality in prehistoric human societies, where moral systems likely first emerged.

3 Agent Cognitive Architecture (MoRE)

Our agent design MoRE is a **m**orality-driven **e**ntity-oriented cognitive processing architecture with **r**eflection capability.

Agent Traits: Moral Types as Example To model evolutionary pressures, agents share a foundational value: maximizing survival and reproduction. It's noteworthy that beyond this, the framework supports researchers to customize agent traits to conduct various research problems in social science beyond this paper's scope. In this paper, we focus on moral simulation and implement moral dispositions based on the "expanding circle" concept (Singer, 1981), defining a morality spectrum:

Self-focused agents care exclusively about themselves. But in implementation, we notice a definitional challenge: purely selfish agents have no interest in reproduction, but defining them to care about offspring would conflate them with kin-focused agents. Therefore, we define them as investing in reproduction but providing no further offspring care, similar to r-selected species (fish, amphibians, invertebrates) that maximize reproductive success through quantity rather than parental care (Pianka, 1970; Stearns, 1992; Gross, 2005; Reznick et al., 2002; Trumbo, 2012)

Kin-focused agents extend moral concern to genetic relatives, providing care and resources to family members while treating non-kin instrumentally.

Group-focused agents extend moral concern beyond kin to include non-related group members. However, defining a group remains a challenge: who constitutes the "group" worthy of moral consideration? We define two variants: 1) *Reciprocal group moral agents* extend care only to those who reciprocate similar moral concern, creating a self-consistent moral circle based on mutual recognition; 2) *Universal group moral agents* extend care to all individuals regardless of their morality. It is expansive, and its non-violent orientation aligns with some intuitive conceptions of morality. However, this variant presents theoretical inconsistencies—violating fairness and reciprocity principles while benefiting agents who may undermine group

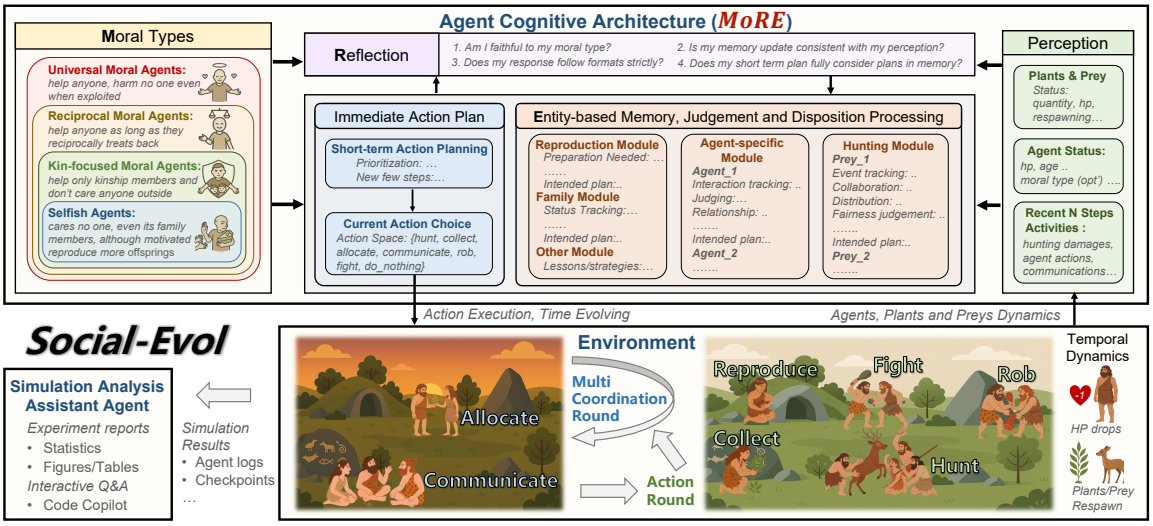


Figure 1: Overview of our simulation framework. The *Environment* follows defined temporal dynamics and iterates between (multi-)coordination rounds and action rounds to interact with agents. The *Agent Cognitive Architecture* includes a moral value module prescribing moral types based on expanding circles of concern, a perception module processing environmental information, and cognitive modules that update memory, form judgments, and generate action plans consistent with the agent's moral type. Before execution, agents perform self-reflection to verify consistency with observed facts and moral dispositions. The *Simulation Analysis Assistant Agent* is a developed tool to automatically analyze simulation results.

welfare, such as selfish agents.

This framework yields four distinct moral types to enable systematic investigation of how different moral dispositions affect evolutionary outcomes. We acknowledge that this discrete categorization simplifies the continuous nature of moral concern in humans for experimental tractability. More importantly, this demonstrates that the framework can customize agent traits by defining the characteristics for focused social science research problems. We note that individual-level agent behaviors are driven by their assigned moral dispositions (i.e., instruction-following). The *emergent* phenomena we study are the population-level dynamics—specifically, which moral type comes to dominate under various ecological constraints—outcomes that are not prescribed by the prompts.

Agent Cognition Framework Our simulation employs agent-based modeling powered by LLM to capture sophisticated cognitive and social dynamics. Agent cognition processes are as follows:

Agent Initialization: Agents are initialized with (1) a profile containing value characteristics aligned with designated moral type; (2) environmental rules governing the simulation; and (3) a knowledge handbook containing a common-sense understanding of environmental dynamics and causal relationships. It ensures agents begin with a comparable baseline understanding without artificially constraining their decision space. Importantly, no agent receives privileged strategic information, preventing methodological bias.

Perception Module: This component processes current status information about plants, prey, and other agents (e.g., HP, age) along with recent activities up to a configurable number of past steps, mirroring human short-term memory.

Cognitive Processing System: We designed an integrated entity-based system that maintains memory, makes judgments, and forms dispositions around entities like other people and hunting animals. This is in contrast to the event-based cognitive processing that records a log-book-like memory and decision history. Our preliminary studies demonstrate that this method effectively prompts LLMs to consider relevant context and perform appropriate reasoning compared to simpler approaches. The entity-based structure provides a template for identifying important information and creating a narrative-like understanding.

Action Planning: This module prioritizes updated memories and dispositional plans to formulate specific actions. This is crucial because the simulation environment may contain many entities toward which an agent might have multiple intended interactions.

Reflection Module: This verification component

ensures cognitive processing and action planning remain consistent with factual information and faithful to the agent's moral type, while producing properly formatted responses.

4 SOCIAL-EVOL Framework

SOCIAL-EVOL (Fig. 1) is a text-based prehistoric hunter-gatherer society with resource constraints, social dynamics, and environmental challenges, enabling research on evolutionary pressures that likely shaped early human morality.

Survival Mechanics Agents maintain health points (HP) that must remain above zero to sustain life. HP diminishes gradually over time, representing basic metabolic costs, and decreases substantially from injuries sustained during hunting or conflict. Agents replenish HP by collecting plants or hunting prey, with successful actions immediately increasing their HP. Additionally, agents face a maximum lifespan constraint that eventually results in death regardless of HP management, modeling natural senescence.

Resource Setting The environment contains two primary resource types: plants and animals. Plants represent low-risk, low-reward resources that agents can reliably collect. Animals offer high-risk, high-reward resources that yield more HP but present significant acquisition challenges. Plants are configured with initial quantity, capacity, nutrition value, and respawn delay. Animals are configured with HP, physical ability, and respawn interval. Those settings can be used to configure the abundance of the available resources.

Production and Reproduction Mechanics

Hunt, Collect: Agents collect plants solely to gain HP without risks whenever plants are available. Hunting allows collaboration, which embodies realistic risk-reward tradeoffs: success probabilities and damage inflicted depend on agent-prey physical ability differentials. Failed hunts cause automatic counter-attacks (i.e., HP loss) to agents; successful attempts progressively reduce prey health until death. This difficulty incentivizes collaboration, improving success rates and distributing risks.

Reproduce: Agents meeting age and HP thresholds can reproduce, though reproduction imposes significant metabolic costs. Offspring begin life with minimal HP, creating vulnerability that typically requires parental investment (food sharing) to ensure survival. Offspring inherit their parents' moral type deterministically; no mutation or cultural transmission mechanism is implemented in the current version, ensuring clean experimental control over inter-generational selection effects.

Social Interaction Mechanics The environment supports multiple forms of social interaction:

Allocate: Agents can voluntarily transfer HP to others, enabling cooperative behaviors and care-based relationships.

Communicate: Agents can exchange messages with specific individuals or groups simultaneously. This communication system enables coordination, information sharing, and relationship development.

Kinship and Type Recognition: The environment explicitly provides each agent with the identities of its parents and offspring. When moral types are configured as observable, this information is included in each agent's perception. When moral types are hidden, agents must infer others' dispositions from observed behaviors and maintain these inferences in memory.

Rob, Fight: Agents may attempt robbery to steal HP or fight to inflict damage without HP gain. Success probability and outcomes depend on relative physical abilities, with gains proportional to the aggressor's strength. Such actions have slightly elevated HP costs due to their intensity. Unlike hunting, failed aggressive attempts do not trigger automatic retaliation; victims must initiate responses independently. This design enables agents to freely express their moral strategies without artificial constraints.

This environmental design creates a complex adaptive system where survival pressures, resource competition, cooperation opportunities, and communication capabilities interact to influence the differential success of varied moral dispositions. Detailed rule explanations and configuration settings appear in the supplement.

Simulation Operating Cycle The simulation operates as a sequential process where agents and the environment interact in defined steps: 1) *Environment Update*, where the simulation refreshes resource availability, agent status changes, and advances time; 2) *Agent Perception*, where each agent receives observations about the current environmental state and recent activities; 3) *Cognitive Processing*, where agents use their architecture to process perceptions, update memory, form judgments, and develop dispositional plans consistent with their moral type toward different entities (prey or other

agents) or goals (reproduction etc.); 4) *Action Planning*, where agents need to consider their dispositional plans and conditions to prioritize and make specific action plans for the next few steps. Note that each simulation step includes one or more coordination rounds (allocate/communicate) and one action round (reproduce/collect/fight/rob/hunt), and the environment update itself with defined temporal dynamics. The number of coordination rounds per step defines the *social interaction cost*: our baseline allows two coordination rounds, while the high-cost condition restricts this to one, directly limiting agents' ability to negotiate and coordinate before acting; and 5) *Consequence Resolution*, where outcomes of all actions are determined. This cycle repeats continually, enabling emergent complex social behaviors while maintaining tractable simulation parameters. The LLM serves as the cognitive engine for each agent, providing reasoning capabilities necessary to navigate moral dilemmas, form social strategies, and respond to environments in ways that reflect human-like cognitive processes.

Two Supported Game Modes Our environment allows two simulation game modes to study both long-term evolution and specific decision dynamics under targeted scenarios. (1) *Evolutionary Game:* The evolutionary games are full simulations of agents' behavior until all agents die or reach maximum steps. This game captures the long-term, emergent outcomes of moral evolution. (2) *Mini Games:* This game setting is designed to isolate a crucial step in the causal chain from morality to fitness: the moment an action is chosen. By placing agents in a specific, controlled scenario like moral dilemmas, we can clearly observe how different moral dispositions translate into distinct behaviors, thus illuminating a key mechanism that drives the broader dynamics seen in the full simulation.

Simulation Data Analysis Assistant Throughout our project development, we identified a significant challenge in LLM-based agent simulations: interpreting the vast quantities of generated data. While having rich, multidimensional data offers tremendous analytical potential, extracting meaningful insights from this complexity requires specialized methodological approaches. To address this challenge, we developed a simulation analysis assistant agent that serves two critical functions. First, it automatically generates comprehensive statistical reports containing the key metrics visualized in our figures. Second, we implemented a series of function calls to enable an interactive Q&A ability when user uses a readily available code copilot agent like Copilot or Cursor. It can allow researchers to interrogate specific agent behaviors, motivations, and decision processes (e.g., "Why did Agent X perform action Y?"). This analytical tool has proven invaluable for understanding simulation dynamics and iteratively refining our agent design architecture. We provide detailed specifications of this system in the supplement.

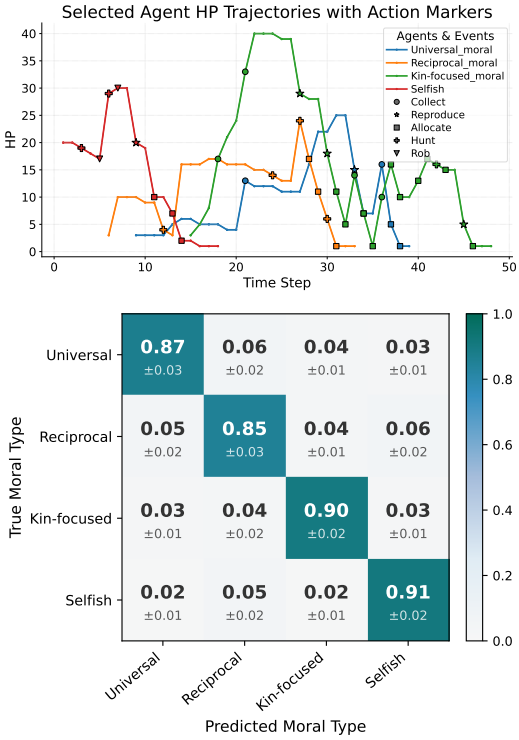


Figure 2: Validation results of the baseline simulation. Upper: HP trajectories of selected agents with action markers. Lower: confusion matrix of moral type inference averaged over 8 runs.

Sim. Model	CM Diag. Acc.	Dom. Type	Final Pop.
GPT-5-mini	0.89 ± 0.03	Kin (6/8)	12.0 ± 2.0
Qwen-3.5	0.86 ± 0.03	Kin (7/8)	11.6 ± 1.7
Kimi-K2.5	0.87 ± 0.03	Kin (5/8)	12.1 ± 1.8

Table 2: Cross-model robustness of the baseline simulation over 8 independent runs per model. CM Diag. Acc. = confusion-matrix diagonal accuracy.

5 Experiments

5.1 Simulation Validation

All experiments use GPT-5-mini as the primary simulation model. To verify that our findings are not artifacts of a specific LLM backbone, we ad-

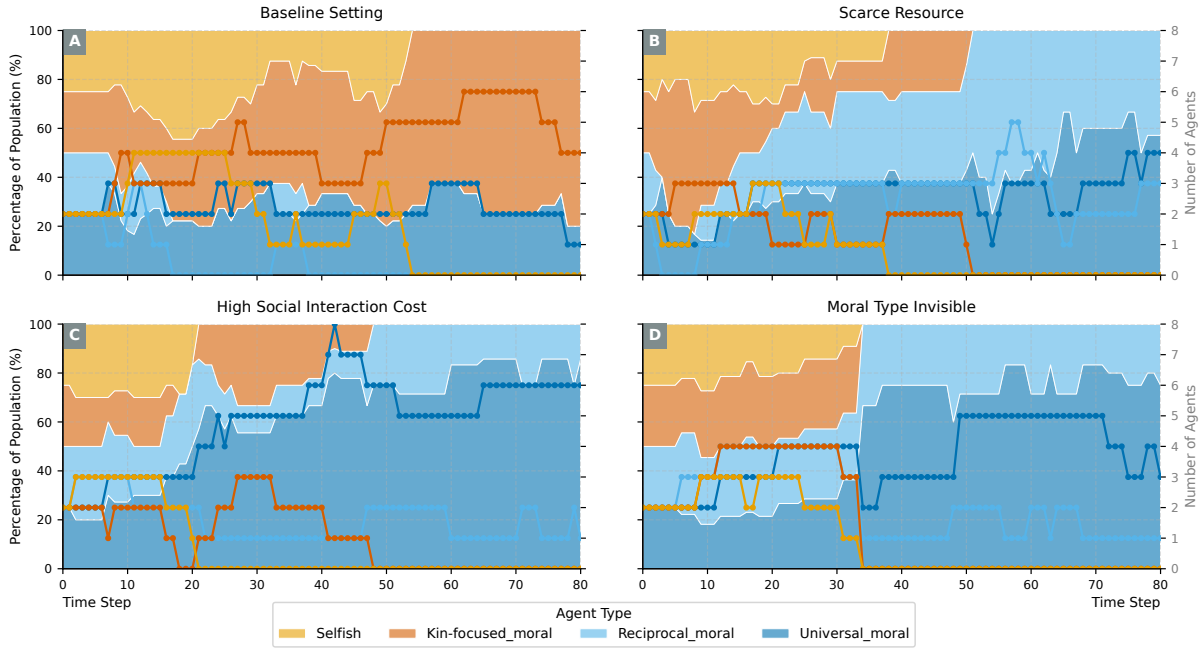


Figure 3: Population dynamics across four experimental settings. Each panel shows a representative run in which the most frequently surviving moral type in that setting persists to the end. (A) Baseline: abundant resources, low social cost, observable moral types. (B) Scarce Resource: reduced resource abundance. (C) High Social Cost: only 1 social round before production. (D) Moral Type Invisible: agents cannot see others' moral types. Outcomes vary across runs; aggregate survival counts over 20 replicate runs are reported in Table 4.

Condition	**CM Diag. Acc.**	Δ
Full Architecture	0.89 ± 0.03	—
Architectural Ablation		
w/o Memory	0.78 ± 0.04	-0.11
w/o Plan	0.82 ± 0.04	-0.07
w/o Reflection	0.84 ± 0.03	-0.05
ReAct Baseline	0.67 ± 0.06	-0.22
Prompt Sensitivity		
Variant A (lexical rewrite)	0.87 ± 0.03	-0.02
Variant B (structural rewrite)	0.86 ± 0.04	-0.03

Table 3: Architecture ablation and prompt sensitivity analysis. All experiments use the baseline configuration with GPT-5-mini.

Setting	**N**	**Uni.**	**Rec.**	**Kin**	**Sel.**
Baseline	8	4	2	**6**	2
Scarce Resource	4	2	**3**	0	1
High Social Cost	4	2	**3**	0	1
Moral Type Invisible	4	**4**	2	2	0

Table 4: Run-level survival counts. A moral type earns one point in a run if it has a nonzero population at step 80; coexistence credits every surviving type.

ditionally run the baseline setting on two other models—the open-source Qwen-3.5 and Kimi-K2.5—and confirm consistent results across all three (Table 2).

We validate from both environmental feedback and morality-behavior consistency perspectives. Fig. 2 (upper) shows HP trajectories for representative agents—e.g., a kin-focused agent that repeatedly sacrifices HP for offspring until it cannot recover. To verify morality-behavior consistency, we use GPT-5 as an evaluator to infer moral types from observed behaviors. Fig. 2 (lower) shows the confusion matrix averaged over 8 runs, achieving a diagonal accuracy of 0.89 ± 0.03. As shown in Table 2, this alignment is consistent across all tested models.

To assess the contribution of each cognitive module in MoRE and the sensitivity to prompt wording, we conduct controlled ablation and prompt-variation experiments (Table 3). Removing any single module degrades behavior-morality alignment, with long-term memory showing the largest impact (-0.11). Stripping all modules (ReAct baseline) reduces diagonal accuracy to 0.67, confirming that each component contributes meaningfully beyond standard LLM reasoning. Prompt sensitivity tests using two semantically equivalent rewrites show negligible variation (≤ 0.03).

5.2 Evolutionary Games

We run experiments across four settings to investigate how moral dispositions interact with environmental and cognitive factors. To characterise stochastic variability, we execute the Baseline $N{=}8$ times and each ablation $N{=}4$ times (20 runs total) with distinct random seeds. Table 4 summarises survival counts; population trajectories for all runs appear in the supplement (§G.2.9). Each simulation is initialized with 8 agents (2 per moral type); representative runs are shown in Fig. 3.

Baseline (Kin: 6/8, Uni: 4/8)**:** We configure abundant resources, sufficient social interaction rounds, and observable moral types. We find that kin-focused agents survive most frequently. We observe that in early stages, universal agents' unconditional helpfulness benefits kin agents, who also participate in broader collaboration when interaction bandwidth permits. However, we find that kin agents partially free-ride on group efforts while channeling surplus resources into family reproduction. Once a kin lineage grows large enough, we observe it becoming self-sustaining through internal family cooperation, gaining a compounding advantage. Universal agents also survive frequently (4/8 runs), indicating that broader moral circles remain competitive even under these favourable conditions.

Scarce Resource (Rec: 3/4, Uni: 2/4)**:** We reduce resource abundance to create scarcity. We find that agents become cautious about cooperation—since moral types are visible, they selectively avoid collaborating with less cooperative types. Reciprocal agents are favoured because their conditional cooperation ensures fair resource sharing within coalitions while excluding free-riders. Notably, we observe rare inter-agent killing behaviours in this setting: selfish agents attack each other first, recognising same-type peers as direct competitors for limited resources. We also observe kin agents targeting selfish agents preemptively. Through mutual aggression and inability to form coalitions, both selfish and kin agents rarely survive. Universal and reciprocal agents persist through coordinated coalition hunting, with reciprocal agents faring slightly better due to their selective cooperation strategy.

High Social Cost (Rec: 3/4, Uni: 2/4)**:** We allow only 1 social round before production. We find that kin-focused agents struggle because they typically require explicit teaming signals—a confirmed commitment from potential partners—before investing effort in collaborative hunts. Under constrained bandwidth, they often cannot receive such signals and therefore avoid risky cooperative ventures, fearing unreturned investment. We observe that universal and reciprocal agents cooperate more naturally: universal agents contribute unconditionally, while reciprocal agents can directly identify other universal and reciprocal agents as trustworthy partners, forming effective coalitions without lengthy negotiation. This leads to higher hunting success rates for both types. Kin and selfish agents, unable to establish effective collaborations, fail to persist.

Moral Type Invisible (Uni: 4/4, Rec: 2/4, Kin: 2/4)**:** We hide moral type labels from all agents. We find that universal agents survive in every run because their cooperation strategy is independent of others' types—they cooperate with everyone regardless, and are therefore never at risk of being misidentified as uncooperative. We observe that reciprocal agents are affected by the noise in type inference: without direct observation, they sometimes misjudge cooperative agents as selfish (or vice versa), narrowing their effective cooperation space. Kin agents also gain some advantage in this setting: since others cannot see that kin agents' helpfulness is directed exclusively toward family, they appear cooperative until extended interaction reveals their selectivity—this delayed identification gives kin agents more time to build family strength. In contrast, selfish agents are quickly exposed through their aggressive and non-cooperative behaviours, leading to their elimination in every run. Detailed mechanistic case studies are provided in the Appendix (§G.3).

5.3 Mini-Games

Beyond long-horizon evolutionary games, we design controlled mini-games to isolate specific mechanisms linking morality to fitness. As an example, we study how moral dispositions shape intergenerational resource sharing within families. We model parent-child dyads across two life stages (young parents with infants, elderly parents with adult children) and four moral dispositions, creating eight scenarios. As shown in Fig. 4, provisioning strategies are systematically driven by moral type: reproductively selfish parents hoard resources for self-preservation, while kin-focused parents consistently prioritize offspring—in extreme cases sacrificing nearly all HP. More mini-game examples are provided in the Appendix.

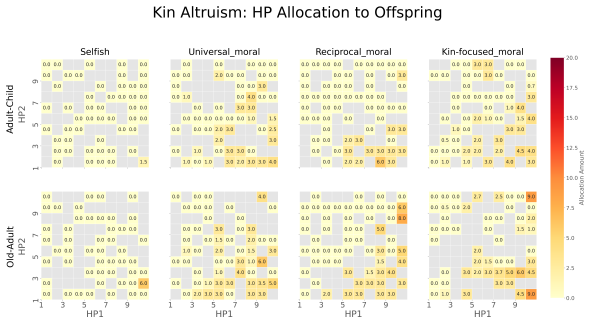


Figure 4: Kin Altruism: A Minigame on Family Resource Allocation.

6 Discussion

Methodological Implications Traditional approaches to studying moral evolution abstract away cognitive processes, encoding behaviour as fixed strategies. Our LLM-based paradigm relaxes these assumptions: agents perceive, remember, reason, and decide through cognitively realistic architectures. This enables cognitive factors (moral type observability, communication bandwidth) to become first-class experimental variables, and provides action-level resolution—tracing *why* a type fails by examining individual agents' reasoning (supplement §G.3).

Cooperation and Mutual Help as the Central Driver Across all experimental settings, our simulations consistently show that *cooperation and mutual help are the primary drivers of evolutionary survival* (Table 4). Universal and reciprocal agents—the most cooperative types—exhibit the most stable survival rates across all conditions. Kin-focused agents remain viable when resources and communication bandwidth suffice for them to reach self-sustaining family size. Selfish agents are strongly disfavoured in every setting. These results are broadly consistent with prior theoretical predictions (Hamilton, 1964; Trivers, 1971; Nowak and Sigmund, 1998; Curry et al., 2019b), and support the expanding circle model (Singer, 1981): broader moral circles generally produce superior evolutionary outcomes.

Emergent Cognitive Mechanisms Beyond these core results, two central mechanisms *emerge from cognition*—neither is explicitly encoded, yet both decisively shape evolutionary outcomes and deepen our understanding of why cooperation prevails.

Self-purging of selfish agents. Because cognitively realistic selfish agents reason about competitors, they recognise same-type peers as threats and preemptively attack them. Combined with exclusion by moral coalitions, selfishness fails not only from lacking cooperation but from actively generating internal conflict—a direct consequence of cognition applied to a selfish value system. Cooperation thus prevails through both its positive benefits *and* the self-destructive dynamics of its absence.

Cost of moral judgment. Assessing others' trustworthiness takes time under limited lifespans and carries the risk of misjudgment—wrongly excluding allies or being wrongly excluded. Universal agents benefit uniquely: they never produce behaviours that could be misread as hostile, so their reputation settles quickly. When judgment cost is high, a "willing-to-take-a-loss" strategy builds reputation faster and secures better cooperation—connecting to costly signalling (Gintis et al., 2001; Nowak and Sigmund, 2005) and bounded rationality (Simon, 1991). This mechanism explains why universal morality is the most consistently viable type across challenging conditions: its unconditional cooperation minimises the cognitive overhead that other strategies must bear.

These patterns resonate with real societies—purely selfish individuals are rare; family-oriented and community-oriented dispositions coexist—and with anthropological evidence that kin-focused cooperation in prehistoric societies (Holden and Mace, 2003; Mattison et al., 2011) preceded the expansion of moral circles under increasing environmental pressures (Singer, 1981). Further emergent phenomena and theoretical connections are detailed in the Appendix.

Conclusions We propose a new research paradigm that brings cognitive realism to the study of moral evolution through LLM-based agent simulation. Our experiments show that cooperation and mutual help are the central driver of evolutionary survival, with universal and reciprocal morality exhibiting the most stable outcomes. Two mechanisms emerge from cognition—self-purging of selfish agents and the cost of moral judgment—that decisively shape evolutionary outcomes and are inaccessible to traditional approaches. We hope this paradigm can inspire broader and deeper investigations into moral evolution and other complex social phenomena that require cognitively realistic modeling.

Limitations

This work is constrained by several design choices that delimit the scope and generality of the presented results.

First, the framework currently relies on general-purpose LLMs for key reasoning and decision-making functions. In preliminary experiments, these models exhibit brittleness in fine-grained spatial and temporal computations, which can propagate into downstream behaviors and reduce the fidelity of simulated interaction dynamics. Addressing such reasoning failures will likely require dedicated mechanisms (e.g., specialized representations or constrained inference procedures) beyond the present implementation.

Second, the current model does not explicitly implement sexual selection, despite its central role in evolutionary dynamics. As a consequence, the framework cannot capture phenomena such as mate choice, mating competition, or selection pressures arising from reproductive strategies. Incorporating these processes would require additional, carefully specified mechanisms and is therefore left to future work.

Third, our environment is intentionally limited to a hunter-gatherer setting. Important components of more complex societies—such as tool innovation and accumulation, exchange/market dynamics, institutional governance, and technologically mediated coordination—are not modeled. While these additions could increase ecological realism, they may also introduce additional confounds and complicate experimental control; thus, conclusions drawn from the present experiments should be interpreted as pertaining to simplified social-ecological conditions.

Ethical considerations

Our project is, at its core, a simulation study of ethics itself. As such, it does not raise the typical ethical concerns associated with methodological research that might be misused. Importantly, our findings can be interpreted as supporting the general proposition that morality is beneficial for humans. The factors that sometimes cause moral agents to fail in evolutionary competition can, in fact, offer valuable insights for promoting social causes and designing mechanisms to enhance the evolutionary advantage of moral individuals. However, we caution against the simplistic interpretation that the conditions under which moral agents fail to prevail are evidence that morality is not advantageous for humans. Such a view is an oversimplification. First, modern society differs profoundly from prehistoric hunter-gatherer contexts. Humans have evolved to be born with moral dispositions (Hamlin et al., 2011; Bloom, 2013; Warneken and Tomasello, 2006). Also in contemporary human societies, almost no one can survive without collaboration, promoting moral behaviors. Second, it is crucial to understand the specific causal role that morality plays in success or failure. For example, our results show that when communication is prohibitively costly, moral agents may be outcompeted by selfish ones. This occurs because morality often inclines agents toward collaboration, which may not be optimal in situations where cooperation is particularly costly. However, morality does not require agents to cooperate indiscriminately; moral agents could, in principle, maintain their moral disposition while choosing to act independently when cooperation is not advantageous, and then collaborate when conditions improve. As revealed by our simulation and common wisdom, being moral does not guarantee success in every circumstance, but a lack of morality fundamentally constrains one's potential for success.

Use of AI Assistants

This paper was primarily conceived, designed, and drafted by the human authors. AI assistants (including ChatGPT and Claude) were used in a supporting role for proofreading, rewriting for clarity, and assisting with code development for the simulation platform. All scientific contributions, experimental design, analysis, and intellectual direction were driven by the authors, with AI tools serving as aids for language refinement and coding assistance.

References

Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. *arXiv preprint arXiv:2306.07872*.

Solomon E Asch. 1946. Forming impressions of personality. *The Journal of Abnormal and Social Psychology*, 41(3):258.

Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *Science*, 211(4489):1390–1396.

Albert Bandura. 1977. *Social learning theory*. Prentice Hall.

Paul Bloom. 2013. *Just babies: The origins of good and evil*. Crown.

Samuel Bowles and Herbert Gintis. 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical population biology*, 65(1):17–28.

Robert Boyd, Peter J Richerson, and Joseph Henrich. 2011. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10918–10925.

Oliver Scott Curry, Daniel A Mullins, and Harvey Whitehouse. 2019a. Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, 60(1):47–69.

Oliver Scott Curry, Daniel Austin Mullins, and Harvey Whitehouse. 2019b. Is it good to cooperate? testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, 60(1):47–69.

Gordon Dai, Weijia Zhang, Jinhan Li, Siqi Yang, Srihas Rao, Arthur Caetano, Misha Sra, and 1 others. 2024. Artificial leviathan: Exploring social evolution of llm agents through the lens of hobbesian social contract theory. *arXiv preprint arXiv:2406.14373*.

Richard Dawkins. 1976. *The selfish gene*. Oxford University Press.

Morton Deutsch. 1973. *The resolution of conflict: Constructive and destructive processes*. Yale University Press.

Ernst Fehr and Simon Gächter. 2002. Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Hsiao-Tung Fei. 1992. *From the Soil, the Foundations of Chinese Society*. University of California Press, Berkeley. Translated by Gary G. Hamilton and Wang Zheng.

Alan P Fiske. 1992. The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological Review*, 99(4):689–723.

Carol Gilligan. 1982. *In a different voice: Psychological theory and women's development*. Harvard University Press.

Herbert Gintis, Eric Alden Smith, and Samuel Bowles. 2001. Costly signaling and cooperation. *Journal of theoretical biology*, 213(1):103–119.

Kurt Gray, Liane Young, and Adam Waytz. 2012. Moral judgment is not a moral science. *Psychological Review*, 119(3):542–572.

Joshua D Greene. 2013. *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Mart R Gross. 2005. The evolution of parental care. *The Quarterly Review of Biology*, 80(1):37–45.

Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. 2024. Richelieu: Self-evolving llm-based agents for ai diplomacy. *Advances in Neural Information Processing Systems*, 37:123471–123497.

Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

Jonathan Haidt. 2007. The new synthesis in moral psychology. *Science*, 316(5827):998–1002.

Jonathan Haidt and Craig Joseph. 2007. The moral mind: How 5 sets of innate moral intuitions guide the development of many culture-specific virtues, and perhaps even modules. In Peter Carruthers, Stephen Laurence, and Stephen Stich, editors, *The innate mind: Volume 3: Foundations and the future*, pages 367–391. Oxford University Press.

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

J Kiley Hamlin, Karen Wynn, and Paul Bloom. 2011. Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26(1):30–39.

Joseph Henrich. 2015. *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Clare Janaki Holden and Ruth Mace. 2003. Spread of cattle led to the loss of matrilineal descent in africa: a coevolutionary analysis. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1532):2425–2433.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

Yizhe Huang, Xingbo Wang, Hao Liu, Fanqi Kong, Aoyang Qin, Min Tang, Xiaoxi Wang, Song-Chun Zhu, Mingjie Bi, Siyuan Qi, and 1 others. 2024. Adasociety: An adaptive environment with social structures for multi-agent decision-making. *Advances in Neural Information Processing Systems*, 37:35388–35413.

Hillard Kaplan, Kim Hill, Jane Lancaster, and A Magdalena Hurtado. 1992. The evolution of life history theory: A bibliometric study of an interdisciplinary research area. *Evolutionary Anthropology: Issues, News, and Reviews*, 1(2):62–71.

James Konow. 2003. Which is the fairest one of all? a positive analysis of justice theories. *Journal of Economic Literature*, 41(4):1188–1239.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Siobhán M Mattison, Eric Alden Smith, Mary K Shenk, and Ethan E Cochrane. 2011. The evolutionary ecology of despotism. *Evolution and Human Behavior*, 32(5):334–347.

Martin A Nowak. 2006. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563.

Martin A Nowak and Karl Sigmund. 1998. The dynamics of indirect reciprocity. *Journal of theoretical biology*, 194(4):561–574.

Martin A Nowak and Karl Sigmund. 2005. Evolution of indirect reciprocity. *Nature*, 437(7063):1291–1298.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Eric R Pianka. 1970. On r and k selection. *The American Naturalist*, 104(940):592–597.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, and 1 others. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Tage S Rai and Alan P Fiske. 2011. Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, 118(1):57–75.

David Reznick, Michael J Bryant, and Farrah Bashey. 2002. r-and k-selection revisited: the role of population regulation in life-history evolution. *Ecology*, 83(6):1509–1520.

Paul Rozin, Laura Lowery, Sumio Imada, and Jonathan Haidt. 1999. The cad triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, 76(4):574–586.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Herbert A Simon. 1991. Bounded rationality and organizational learning. *Organization science*, 2(1):125–134.

Peter Singer. 1981. *The expanding circle: Ethics, evolution, and moral progress*. Princeton University Press.

Stephen C Stearns. 1992. *The Evolution of Life Histories*. Oxford University Press, Oxford.

Henri Tajfel. 1979. Individuals and groups in social psychology. *British Journal of Social and Clinical Psychology*, 18(2):183–190.

Michael Tomasello. 2016. *A natural history of human morality*. Harvard University Press.

Robert L Trivers. 1971. The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1):35–57.

Stephen T Trumbo. 2012. Patterns of parental care in invertebrates. In Nick J Royle, Per T Smiseth, and Mathias Kölliker, editors, *Evolution of Parental Care*, pages 81–100. Oxford University Press, Oxford.

Elliot Turiel. 1983. *The development of social knowledge: Morality and convention*. Cambridge University Press.

Yiding Wang, Yuxuan Chen, Fangwei Zhong, Long Ma, and Yizhou Wang. 2024. Simulating human-like daily activities with desire-driven autonomy. *arXiv preprint arXiv:2412.06435*.

Zhilin Wang, Yu Ying Chiu, and Yu Cheung Chiu. 2023. Humanoid agents: Platform for simulating human-like generative agents. *arXiv preprint arXiv:2310.05418*.

Felix Warneken and Michael Tomasello. 2006. Altruistic helping in human infants and young chimpanzees. *Science*, 311(5765):1301–1303.

David Sloan Wilson and Edward O Wilson. 2007. Rethinking the theoretical foundation of sociobiology. *The Quarterly Review of Biology*, 82(4):327–348.

A General Discussions

A.1 Code Release

Our code has been released under the MIT license at https://github.com/MoralAgentSim/Simulation-Engine. A project page with documentation and examples is available at https://MoralAgentSim.github.io. This platform will be actively maintained and updated to support more features and research questions. We welcome any collaboration, contribution, feedback, and feature requests.

A.2 Potential Risks

Despite our efforts toward realism, our simulation operates in a constrained environment that inevitably omits many real-world factors. Results should therefore be treated as generating insights and hypotheses rather than definitive conclusions, and should not be used to directly guide important cognitive or policy decisions. Additionally, findings about conditions under which selfish strategies occasionally persist could, in principle, be interpreted as guidance for when self-interested behaviour might go unchecked—though the high level of abstraction in our simulation limits such applicability.

A.3 More Discussion Over Our Methodology

As we have emphasized, our method should be viewed as a complement to traditional mathematical models, not a replacement. By incorporating rich psychological realism into the simulation, our approach enables researchers to investigate how numerous factors interact in complex ways. However, this increased realism also means that simulation results are sensitive to the specific details of these factors and may not yield the definitive answers that highly abstract mathematical models can provide.

Yet, definitive answers are not always the primary goal of research, especially in the social sciences. Often, the objective is to uncover previously unnoticed factors that influence a phenomenon or to explore the intricate interplay among multiple variables. Such goals are difficult to achieve with traditional mathematical models, which require all relevant factors to be known or assumed in advance. Historically, researchers have relied on field studies to observe human behavior and identify these factors, but simulation now offers a cost-effective means to assist in discovery and hypothesis generation, potentially accelerating progress in the field.

Moreover, when the number of interacting factors becomes too great for analytical calculation, simulation becomes indispensable. While simulations inevitably deviate from reality—just as any modeling method, and such deviations may be amplified in large-scale runs—they can still provide valuable insights into research questions. Simulations can reveal what is possible, and the underlying mechanisms and developmental dynamics they expose may remain relevant even if the precise outcomes differ from those observed in the real world.

A.4 Flexibility of the Simulation Platform

Our platform is designed to be flexible and extensible. By varying the configuration settings, one can use the same platform to study different research questions. For example, we have used the same platform to study the effect of different moral types, different resource distributions, different communication costs, etc. In the section below, we also list a list of findings that are connected to different research areas that could possibly be investigated further with our platform.

Moreover, we support researchers to extend beyond morally related value dispositions. One can flexibly define the value dispositions of the agents by writing appropriate prompt templates. For example, one can define agents to be of different cultural backgrounds, different religions, different political views, etc. Or one can also study the effect of specific social norms by prescribing the agents to follow certain social rules, e.g, always equal distribution VS always contribution-based distribution, etc. Hunting-gathering environment equipped with general social interaction dynamics is very general to support a wide range of research questions.

B Discovered Phenomena That Connect to Other Theories

As mentioned, one key feature of what our platform can provide is that we can naturally see a lot of emergent phenomena that matter for social evolution regarding morality. These phenomena were abstracted away in the traditional mathematical models. But on our platform, they will surface on their own to deepen our understanding. Those phenomena or topics were traditionally a subject of research areas on their own, but now we can study them in a unified framework.

We list some of the observed phenomena and identify some of the theories that are related to them in Table 5. This list is definitely not exhaustive. We hope this can provide a good starting point for future researchers to discover more phenomena and theories.

We also encourage researchers to use our platform as a new way to study these phenomena and theories.

C System Design Details

C.1 Simulation Pipeline

The general system workflow functions as Figure 5. System first initializes the environment based on the system setting config (e.g, see Table 13) or resumes from the previous experiment run. The specific initialization phases are shown in Table 6.

Then the system enters into an execution cycle that allows agents to perceive and perform cognitive processing to plan for actions and update the environment accordingly. The execution phases are shown in Table 7. Within this cycle, there is also a system validation and correction cycle over the agent's response and action to ensure its format and content are legal (see Figure 6 and Table 8).

Please refer to those tables and figures for more details.

Table 5: Discovered Phenomena and Related Theories

Phenomena Findings from Experiments	**Related Theories**
Coordination is costly: • Communication takes time and can reduce the time for other important things.	*Coordination Cost Theory* (Simon, 1991) "Organizations face bounded rationality where coordination costs limit optimal decision-making"
Moral judgment based on actions: • Agents evaluate others' morality by observing how they treat third parties. • Actions toward others, not just toward oneself, shape moral reputation.	*Moral Judgment Theory* (Haidt, 2001) "People make rapid moral judgments based on observed behaviors and their emotional responses" *Impression Formation* (Asch, 1946) "Observers form impressions of others' character based on their actions toward third parties"
Misunderstandings can lead to major conflicts: • Agents may misinterpret others' intentions or actions, leading to unnecessary conflicts. • Limited communication can cause agents to make incorrect assumptions about others' moral types or goals.	*Communication Theory* (Shannon, 1948) "Information transmission is inherently imperfect, leading to potential misunderstandings and conflicts" *Conflict Resolution* (Deutsch, 1973) "Many conflicts arise from misperceptions and misunderstandings rather than actual incompatible goals"
Predictable morality acts as a reputation/clear signal: • Universal moral agents, by rejecting violence completely, benefit from being clearly understood. • Such behavior is costly though, making them subject to exploitation	*Costly Signaling Theory* (Gintis et al., 2001) "Consistently moral behavior, despite its costs, serves as an honest signal of cooperative intent, reducing the risk of being misjudged as a threat." *Indirect Reciprocity* (Nowak and Sigmund, 2005) "A clear reputation for cooperation, built through predictable actions, is essential for reciprocal altruism and protects an agent from being wrongly punished."
Universal moral agents get exploited: • Agents who never retaliate or punish others' bad behavior become targets of exploitation. • Their unconditional cooperation makes them vulnerable to free-riders.	*Altruistic Punishment* (Fehr and Gächter, 2002) "Cooperation requires punishment of defectors; pure altruism without retaliation is vulnerable to exploitation" *Strong Reciprocity* (Bowles and Gintis, 2004) "Evolutionary success requires both cooperation and punishment of non-cooperators"
Group membership is contested: • Agents might not agree on who is in the group that can share resources.	*Social Identity Theory* (Tajfel, 1979) "Group boundaries are fluid and contested, with membership determined by shared identity markers and mutual recognition"
Distribution methods are complex: • How to distribute? Distribute evenly, based on contribution, harm taken, need, can affect both the success of the end result and each other's judgment.	*Distributive Justice* (Konow, 2003) "Fairness judgments depend on multiple principles, including equality, need, and contribution"
Careful planning is important: • Reproduction schedule is important. Too frequent can cause both parents and children to die.	*Life History Theory* (Kaplan et al., 1992) "Organisms face trade-offs between current and future reproduction, with timing being crucial for survival"
Tendency to cooperate can sometimes have negative effect: • Moral agents have a tendency to collaborate to acquire resources, but in some particular setting (with competition, resources being in some way), taking faster action instead of collaboration may be more crucial. • Moral agents tend to agree to collaborate to hunt, but they might not be in a good position to hunt.	*Cooperation Dilemmas* (Bowles and Gintis, 2004) "Cooperation can be maladaptive when individual action would yield higher returns"
Moral agents' mutual dependency sometimes leads to disaster end: • Moral agents tend to trust others to help them later, but the others may also think the same, and none have the extra capacity to help.	*Trust and Cooperation* (Fehr and Gächter, 2002) "Altruistic punishment can maintain cooperation but may lead to cascading failures when trust is misplaced"
Mutual reinforcing / social pressure: • When some agents reproduce, others feel compelled to do so too, even though their HP was not very high.	*Social Learning Theory* (Bandura, 1977) "Social learning and imitation can lead to behavioral contagion even when not optimal for individuals"

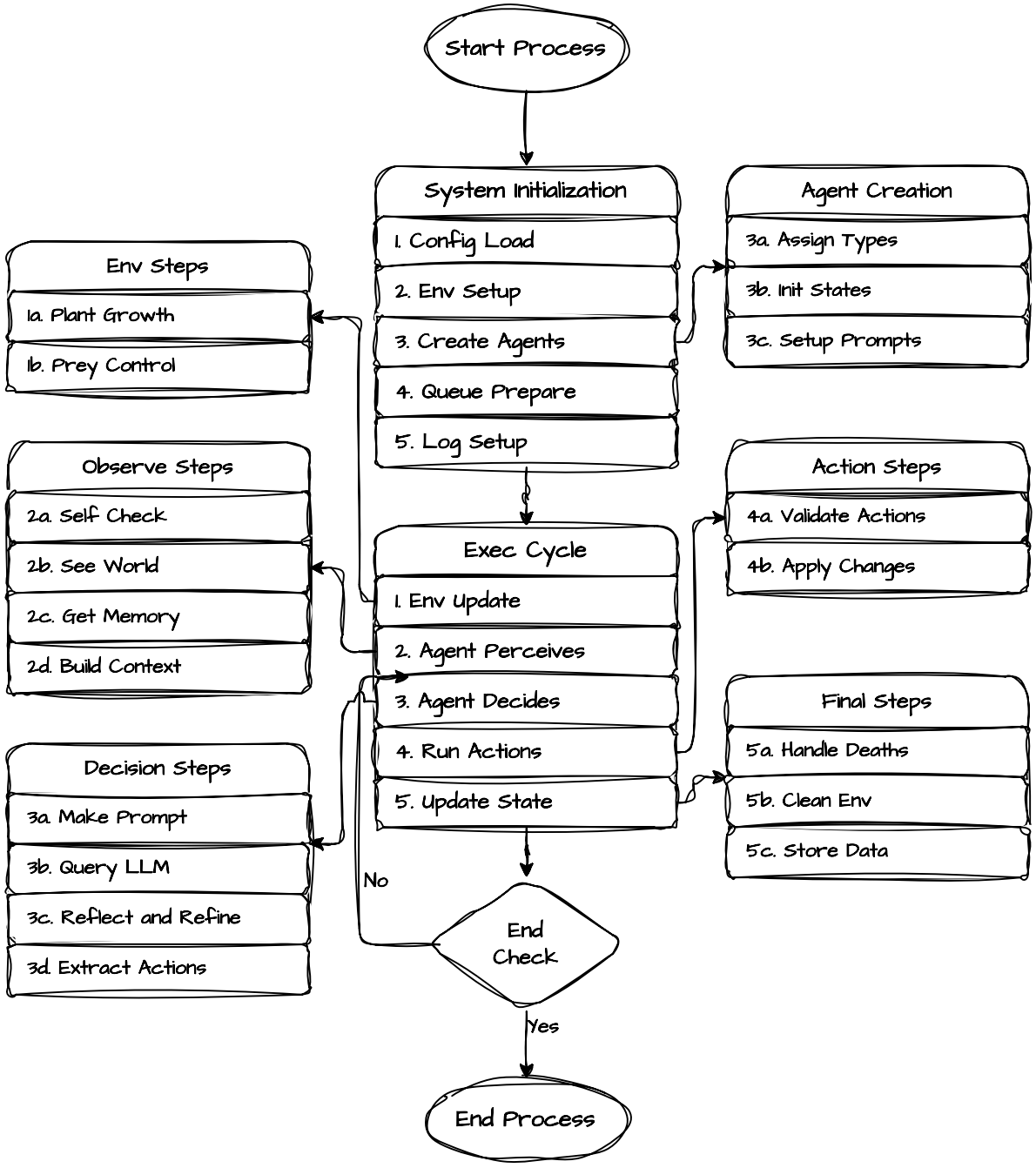


Figure 5: **Simulation Pipeline Overview** showing the main components and data flow through the system architecture. The pipeline illustrates how the Singleton-based Checkpoint, modular microservices, and key simulation processes interact to maintain a consistent state and flow of information.

Table 6: Simulation Initialization Phases

Phase	Description
Configuration Loading & Validation	• Loads parameters from configuration file (prompt paths, agent types, rules, strategies) • Validates type correctness, constraints, and completeness • Creates authoritative configuration object for simulation
Environment Setup	• Plant Resources: - Generated based on configured abundance - Each plant gets a unique ID, initial quantity, capacity, nutrition value, and respawn delay • Prey Animals: - Initialized with unique IDs - HP and max health sampled from a Gaussian distribution - Assigned physical ability values • Resources placed randomly in unoccupied grid cells
Initial Agent Spawning	• Instantiates agents based on population size • Assigns moral types according to configuration ratios • Initializes attributes: HP, age, physical ability • No initial family ties
Execution Queue Setup	• Creates randomized agent sequence for fair execution • Initializes time step counter (typically 0 or 1) • Sets up containers for agent observations
Logging System Setup	• Configures comprehensive tracking system • Creates log files for: - Global progress summaries - Per-step execution records - Detailed event logs - Error diagnostics • Organizes logs in uniquely named directories

Table 7: Per-Step Execution Cycle Phases

Phase	Description
Environment State Update	• Updates plant lifecycle: restores depleted plants after respawn delay, increases quantity for non-depleted plants • Spawns new prey in empty locations based on probability and maximum count • Removes dead prey from the grid
Agent Observation	• Self-assessment: queries HP, age, inventory, physical ability, reproductive status • Environmental perception: detects nearby resources, prey, and other agents • Memory retrieval: accesses past observations, messages, and action outcomes • Context formatting: structures information for LLM prompt
Agent Decision Making	• Constructs system message with agent persona and rules • Builds user message with current state and context • LLM processes context and returns proposed action • Validates response format and structure
Action Execution & Validation	• Performs response & action validation • Applies validated actions to simulation state
State Finalization	• Updates agent HP, inventories, and environmental quantities • Handles communication and memory updates • Performs system-wide consistency checks • Records detailed logs of agent states, environment state, and metrics • Prepares state for next cycle
Termination Check	• Evaluates termination criteria (max steps, population collapse, goals) • Either concludes simulation or increments time step

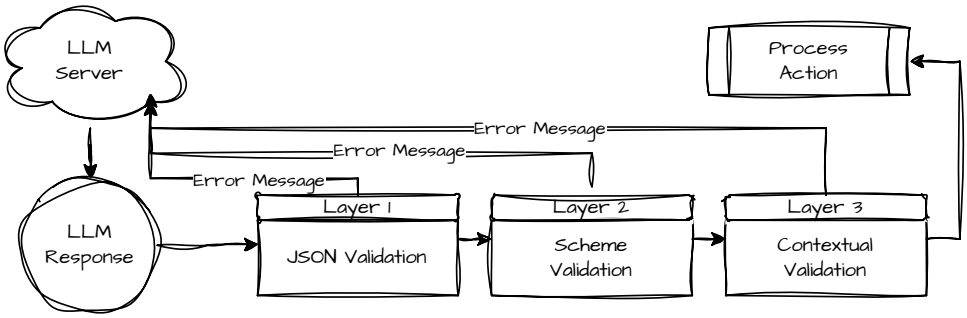


Figure 6: Multi-layer validation and retry framework showing the escalating levels of validation applied to agent actions. The diagram illustrates the three validation layers: syntactic and schema validation, contextual rule-based pre-validation, and action handler final validation, along with their respective feedback loops and retry mechanisms.

Table 8: The Checklist of Multi-Layer Response & Action Validation.

Layer	**Description**
Layer 1: Syntactic and Schema	• Applied immediately after LLM output generation • Two critical checks: - Syntactic validation: Ensures proper JSON formatting - Schema validation: Verifies required fields, types, and enumerations • Retry mechanism: - Resends prompt with error metadata - Limited to predefined maximum attempts • Focus: Structural correctness only
Layer 2: Contextual and Rule-Based	• Domain-specific validation within Agent Decision Making phase • Contextual checks: - Target existence and accessibility - Location-based constraints - HP sufficiency for action costs • Memory constraints: - Long-term memory capacity limits • Feedback loop: - Human-readable error messages - Updated prompts with feedback - Configurable retry rounds
Layer 3: Action Handler Final	• Executed during Action Execution phase • Domain-specific validation in action handlers • Dynamic condition checks: - Agent adjacency for physical interactions - HP sufficiency with current state - Race conditions with shared resources • No LLM retry mechanism • Failure handling: - Action nullification or failure processing - Logging to agent observation history

C.2 System Design Principles

The Morality-AI simulation is built on two core principles: centralized state management and modular architecture. A Singleton-based Checkpoint class maintains a single, authoritative simulation state, ensuring consistency, atomic updates, and easy reproducibility. This design prevents conflicting states and simplifies debugging and resuming experiments. The system adopts a microservice-inspired structure, separating major functions—such as state persistence, agent reasoning, and LLM interfacing—into independent, easily testable modules. This modularity enhances maintainability, scalability, and flexibility, allowing components to be updated or replaced without affecting the overall system. Together, these principles provide a robust and extensible foundation for complex agent-based simulations.

D Agent Design Details

D.1 Agent Designs and Workflows

Agents are the primary decision-making entities in the simulation. They possess a set of core attributes that govern their physical capabilities, cognitive constraints, and eligibility for specific actions (see the agent attributes in Table 13).

At the beginning of agent initialization, agent will be given their value/moral type prompt and all the system prompts like environment dynamics, requirement, commonsense strategies etc (prompt details see F). Then, during each execution cycle, the agent will be given the perception of the environment and its own status, and perform cognitive processing to make action plan. They will perform one round of reflection before finalize their response that contains their cognitive processing and action plan. The process follows Figure 7 to make decisions.

For the current project, the structure of the agent's moral type is listed in Table 9, with the rationale of the design choices in the main text. We want to note that these moral types is not the only way to define the value of an agent. The value can be defined in many other ways - one can focus on the action principles, or calculation of utility, or even be involved in culture and religion to study different problems.

The structure of the agent's perception space is listed in Table 11. The structure of the agent's cognition is listed in Table 10. The content in the action space is listed in Table 12.

D.2 The Quantitative Model for Calculating Action Results

The success rate and damage point in actions like hunting, robbery, etc, is calculated based on the physical ability of the agents. The physical ability values are initialized as a random number from a Gaussian distribution with specified mean and standard deviation in the configuration file Table 13 (with 0 standard deviation, there will be no random variation). Note that prey also has a physical ability value that is initialized in the same way.

Success Rate The success of hunt, fight, and rob actions takes on probabilistic manner. The success of such actions depends on the relative physical abilities of the involved entities. Let $\Delta\text{PA} = \text{PA}_k - \text{PA}_{target}$ represent the physical ability differential between an actor k and a target entity (which could be another agent j or a prey animal A_j). The probability of success, P_{succ}, for these actions is determined by the function:

$$P_{succ}(\Delta\text{PA}; I_{\text{PA},k}, S_{\text{PA},k}) = \min \left(\max \left((0.5 + I_{\text{PA},k}) + 0.4 \cdot \tanh \left(\frac{\Delta\text{PA}}{S_{\text{PA},k}} \right), 0.1 \right), 0.9 \right)$$

Here, $I_{\text{PA},k}$ and $S_{\text{PA},k}$ are agent k's specific scaling parameters (an intercept offset and a slope divisor, respectively) pertinent to physical ability interactions, derived from its configuration. The function $\min(\max(x, a), b)$ ensures the probability is clipped to the interval $[a, b]$, in this case, $[0.1, 0.9]$. The outcome of such an action is then determined by a Bernoulli trial $X \sim \text{Bernoulli}(P_{succ})$.

In the descriptions that follow, $\text{HP}_k(t')$ signifies the health of agent k after any initial action-specific costs have been deducted, but before other consequences of the action (e.g., gains from success, damage from failure) are applied.

Collect Agent k may attempt to gather resources from a designated plant node P_i, which possesses a current resource quantity $Q_i(t)$. The agent specifies a desired quantity q_{req}. For the action to be valid, P_i must be a plant, and its available quantity must meet the request, i.e., $Q_i(t) \geq q_{req}$. The actual quantity gathered, q_{coll}, is constrained by the request, availability, and the agent's single-action collection capacity, $k_{collect}$ (a global limit):

$$q_{coll} = \min(q_{req}, Q_i(t), k_{collect})$$

A positive quantity must be collectible ($q_{coll} > 0$). Consequently, the agent's health and the plant's

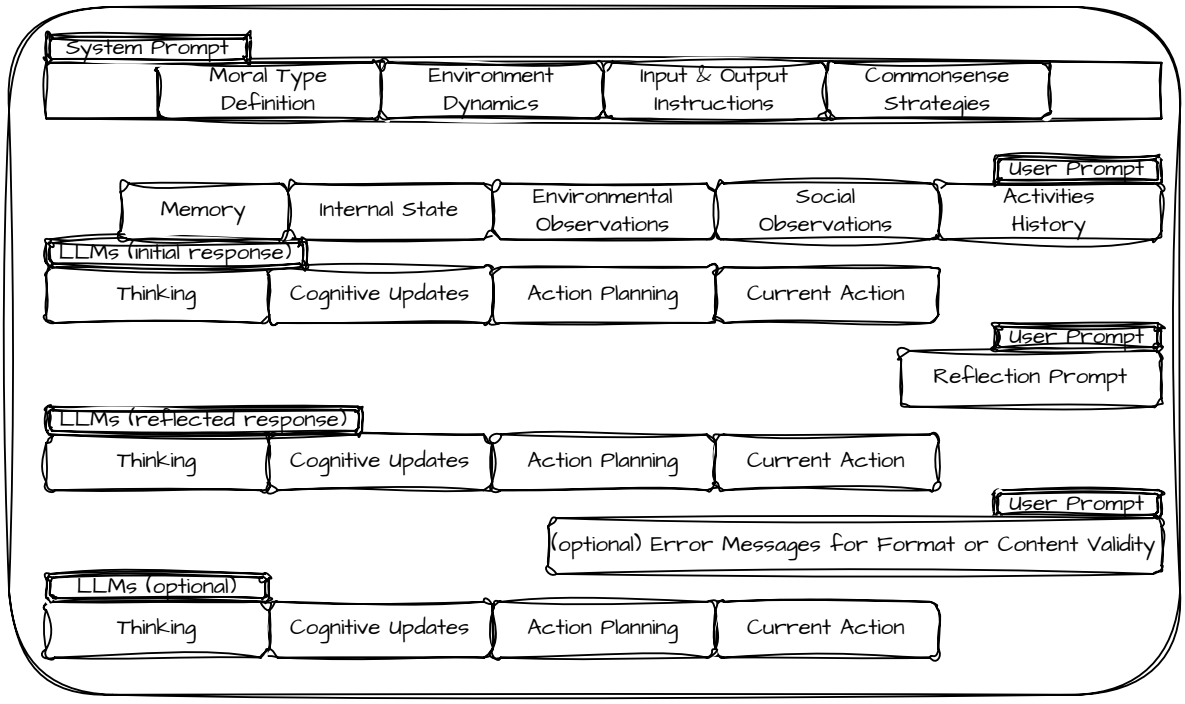


Figure 7: The LLM query process for decision making, illustrating the flow from observation gathering through prompt construction, LLM interaction, and action validation. This process shows how environmental perceptions, agent state, and memory are integrated to produce contextually relevant decisions within the simulation environment.

Table 9: Agent Moral Types Summary. Here, we summarize the core characteristics, expected typical behaviors, and expected cooperation patterns of these moral types. However, the simulated behavior for each agent might not strictly follow the expected behaviors due to the randomness of LLM's output. The exact prompt for each moral type is shown in F

Moral Type	**Core Characteristics**	**Expected Typical Behaviors**	**Expected Cooperation Pattern**
Universal Group-Focused Moral	Aim for universal well-being and collective good, harm-action averse	Share resources freely; protect others from harm; communicate transparently	Highly altruistic and cooperative with all agents
Reciprocal Group-Focused Moral	Fairness and mutual benefit within in-group, harm action allowed	Form strong bonds with cooperative peers	Cooperative with in-group; neutral or adversarial to outgroup or selfish agents
Kin-Focused Moral	Prioritize genetic relatives above all else, harm action allowed	Form close-knit kinship clusters; sacrifice for kin	Intensely altruistic toward family; indifferent or competitive toward non-kin
Reproductive Selfish	Personal reproductive success, harm action allowed	Acquire resources for own survival; opportunistic tactics	Cooperate only when serving reproductive interests; inclined to hoard resources

Table 10: Agent Cognition Structure (Memory, Judgement and Planning)

Field	Key Subfields / Description
1. Prey-Based Cognition	• Organized by prey_id: ○ hunt_fact_history_of_this_prey: who hunted, effect, time step, damage, if_killed ○ communication_and_planning_before_killing_prey: reward, collaborators, distribution plan, objections ○ distribution_after_killing_prey: winner, allocation, fairness evaluation, free rider check ○ plan_next: next plan, retaliation plan, stage, reasoning ○ afterward_happenings: retaliation events, other events, lessons learned
2. Agent-Based Cognition	• Organized by agent_id: ○ important_interaction_history: what_i_did_to_him, what_he_did_to_me (action type, success, reason, target moral type) ○ thinking: evaluation, judgement, relationship, agreement, plan
3. Family Plan	• Organized by agent_id: ○ status: how the family member is doing ○ plan: what to do to/with them
4. Reproduction Plan	• thinking: reasoning about reproduction plan • preconditions_and_subgoals: specific preconditions needed • estimated_time_to_produce_next_child: time step
5. Learned Strategies	• Lessons learned, strategies to follow in the future

Table 11: Agent Perception Content Structure

Category	Description
Self/Internal Information	• Current HP and health status • Family relationships and status • Personal attributes and capabilities
Environment Status	• Available plant resources • Prey animals present in the environment • Resource locations and quantities
Other Agents Status	• Basic information (age, HP) of other agents • Moral types of other agents • Current positions and states
Recent History	• Last 15 steps of personal interactions: - Others' actions and communications toward self - Self's actions toward environment and others • Recent events: - Changes in environment and other agents - Family-related news and updates • Hunting activities: - Personal involvement in prey hunting - Related communications and outcomes
Memory	• Updated memory from previous step • Immediate action plans from previous step

Table 12: Agent Action Space Summary

Action	Description	Requirements	HP Cost	Outcome
(Re)Production Actions				
Collect	Gather plant resources from environment	Resource exists and is a plant node	None	Agent gains HP (quantity $\times$ nutrition value); plant quantity reduced
Hunt	Target prey animals for nutritional gain	Prey exists	1 HP + additional damage if failed	If successful, prey killed and agent receives reward equal to prey's max HP
Reproduce	Create offspring	Minimum age and HP thresholds met	Defined in reproduction parameters	Child agent created with age 0 and initial HP, inheriting parent's archetype
Social Interaction Actions				
Allocate	Transfer HP to other agents	Targets exist and are alive; sufficient HP	Equal to HP transferred	Recipients gain specified HP (capped at maximum)
Fight	Attempt to damage another agent	Target exists, is alive, not self	1 HP resistance cost	If successful (based on ability difference), target suffers damage equal to attacker's ability
Rob	Forcibly transfer HP from another agent	Target exists and is alive	1 HP + potential failure penalty	If successful, HP transferred from target to robber
Communicate	Send messages to other agents	Target agents exist and are alive	None	Message recorded in recipient's memory
Other Actions				
DoNothing	Abstain from all actions	None	None	No changes to agent or environment

resources are updated as follows:

$$\begin{aligned}&\mathrm{HP}_k(t+1) = \\ &\min(\max(\mathrm{HP}_k(t) + q_{coll} \cdot H_{plant}, 0), \mathrm{HP}_{k,max}) \\ &Q_i(t+1) = Q_i(t) - q_{coll}\end{aligned}$$

where H_{plant} denotes the nutritional value conferred per unit of the plant resource. This action imparts no direct HP cost to agent k.

Allocate An agent k (the donor) can transfer Health Points to other agents. This is specified via an `allocation_plan`, $(h_{kj})_{j \in J}$, where $h_{kj} \in \mathbb{R}^+$ is the amount of HP designated for transfer to each target agent j in a non-empty set $J \subset \mathcal{K}(t)$. The total HP intended for allocation by agent k is $H_{alloc,k} = \sum_{j \in J} h_{kj}$. This action is permissible if all target agents $j \in J$ are alive and the donor possesses sufficient HP, specifically $\mathrm{HP}_k(t) > H_{alloc,k}$. If valid, the HP of the involved agents is then adjusted:

$$\begin{aligned}&\forall j \in J, \\ &\mathrm{HP}_j(t+1) = \min(\max(\mathrm{HP}_j(t) + h_{kj}, 0), \mathrm{HP}_{j,max}) \\ &\mathrm{HP}_k(t+1) = \\ &\min(\max(\mathrm{HP}_k(t) - H_{alloc,k}, 0), \mathrm{HP}_{k,max})\end{aligned}$$

Fight Agent k (attacker) may engage agent j (target), provided $k \neq j$ and j is alive. To initiate a fight, the attacker k incurs an immediate cost $C_{fight,init} = 1$ HP:

$$\mathrm{HP}_k(t') = \mathrm{HP}_k(t) - C_{fight,init}$$

If $\mathrm{HP}_k(t') \leq 0$, agent k is removed from $\mathcal{K}(t+1)$. Otherwise, the outcome of the fight is determined by a Bernoulli random variable $X_{fight} \sim \text{Bernoulli}(P_{succ}(\Delta\mathrm{PA}_{kj}; I_{\mathrm{PA},k}, S_{\mathrm{PA},k}))$, where $\Delta\mathrm{PA}_{kj} = \mathrm{PA}_k - \mathrm{PA}_j$. The health point dynamics for both the target and attacker, contingent on the outcome X_{fight}, are:

- If $X_{fight} = 1$ (success): The target's health is reduced, $\mathrm{HP}_j(t+1) = \min(\max(\mathrm{HP}_j(t) - \lfloor \mathrm{PA}_k \rfloor, 0), \mathrm{HP}_{j,max})$.
- If $X_{fight} = 0$ (failure): The target's health remains unchanged, $\mathrm{HP}_j(t+1) = \mathrm{HP}_j(t)$.

In both scenarios, the attacker's health after the interaction resolves is $\mathrm{HP}_k(t+1) = \mathrm{HP}_k(t')$. The target agent j is removed if its health $\mathrm{HP}_j(t+1) \leq 0$.

Rob Agent k (robber) may attempt to forcibly extract $h_{rob,req} > 0$ HP from a target agent j, provided j is alive and possesses sufficient health ($\mathrm{HP}_j(t) \geq h_{rob,req}$). The robber k first incurs an initiation cost $C_{rob,init} = 1$ HP:

$$\mathrm{HP}_k(t') = \mathrm{HP}_k(t) - C_{rob,init}$$

If $\mathrm{HP}_k(t') \leq 0$, k is removed. Otherwise, the success of the attempt is a random variable $X_{rob} \sim \text{Bernoulli}(P_{succ}(\Delta\mathrm{PA}_{kj}; I_{\mathrm{PA},k}, S_{\mathrm{PA},k}))$, with $\Delta\mathrm{PA}_{kj} = \mathrm{PA}_k - \mathrm{PA}_j$. Depending on the outcome X_{rob}, the HP updates are:

- If $X_{rob} = 1$ (success):

$$\begin{aligned}&\mathrm{HP}_j(t+1) = \\ &\quad \min(\max(\mathrm{HP}_j(t) - h_{rob,req}, 0), \mathrm{HP}_{j,max}) \\ &\mathrm{HP}_k(t+1) = \\ &\quad \min(\max(\mathrm{HP}_k(t') + h_{rob,req}, 0), \mathrm{HP}_{k,max})\end{aligned}$$

  The target j is removed if $\mathrm{HP}_j(t+1) \leq 0$.

- If $X_{rob} = 0$ (failure): No HP is transferred, thus $\mathrm{HP}_j(t+1) = \mathrm{HP}_j(t)$, and the robber's health remains $\mathrm{HP}_k(t+1) = \mathrm{HP}_k(t')$.

Hunt Agent k (hunter) may target a prey animal A_j, characterized by physical ability PA_{A_j} and health $\mathrm{HP}_{A_j}(t)$ (with maximum $\mathrm{HP}_{A_j,max}$). The hunter k incurs an initial cost $R_{hunt} = 1$ HP:

$$\mathrm{HP}_k(t') = \mathrm{HP}_k(t) - R_{hunt}$$

If $\mathrm{HP}_k(t') \leq 0$, k is removed. Otherwise, the outcome is governed by $X_{hunt} \sim \text{Bernoulli}(P_{succ}(\Delta\mathrm{PA}_{kA_j}; I_{\mathrm{PA},k}, S_{\mathrm{PA},k}))$, where $\Delta\mathrm{PA}_{kA_j} = \mathrm{PA}_k - \mathrm{PA}_{A_j}$.

- If $X_{hunt} = 1$ (success): The prey A_j sustains damage $D_{A_j} = \lfloor \mathrm{PA}_k \rfloor$, leading to $\mathrm{HP}_{A_j}(t+1) = \max(0, \mathrm{HP}_{A_j}(t) - D_{A_j})$. If this damage proves lethal ($\mathrm{HP}_{A_j}(t+1) \leq 0$), prey A_j is removed, and the hunter k gains HP from the kill:

$$\begin{aligned}&\mathrm{HP}_k(t+1) = \\ &\min(\max(\mathrm{HP}_k(t') + \mathrm{HP}_{A_j,max}, 0), \mathrm{HP}_{k,max})\end{aligned}$$

  If the prey survives the damage, the hunter gains no HP from the hit, so $\mathrm{HP}_k(t+1) = \mathrm{HP}_k(t')$.

- If $X_{hunt} = 0$ (failure): The prey A_j counter-attacks, inflicting D_{prey} damage upon hunter k. This D_{prey} is a characteristic of the prey (e.g., its counter-attack strength). The hunter's health is updated to

$$\mathrm{HP}_k(t+1) = \min(\max(\mathrm{HP}_k(t') - D_{prey}, 0), \mathrm{HP}_{k,max})$$

Hunter k is removed if $\mathrm{HP}_k(t+1) \leq 0$.

Reproduce An agent k may create offspring if it meets age and health criteria: $Age_k(t) \geq Age_{repro,min}$ and $\mathrm{HP}_k(t) \geq \mathrm{HP}_{repro,min}$. Upon successful reproduction, a new agent c is added to the population $\mathcal{K}(t+1)$, initialized with $Age_c(0) = 0$ and health $\mathrm{HP}_c(0) = \mathrm{HP}_{child,init}$. The parent k incurs an HP cost, $\mathrm{HP}_{repro,cost}$, resulting in an updated health:

$$\mathrm{HP}_k(t+1) = \min(\max(\mathrm{HP}_k(t) - \mathrm{HP}_{repro,cost}, 0), \mathrm{HP}_{k,max})$$

Communicate Agent k can send a textual message M, constrained by length ($|M| \leq L_{msg,max}$), to a specified set of recipient agents $J \subset \mathcal{K}(t)$. All recipients must be alive. This action does not directly alter HP.

DoNothing An agent k may elect to perform no explicit action. This choice has no effect on its state or the environment; thus, $\mathrm{HP}_k(t+1) = \mathrm{HP}_k(t)$.

E Simulation Analysis Agent System

The Simulation Analysis Agent System is a comprehensive post-processing analysis framework for the Morality-AI simulation environment. It is engineered to distill actionable insights and facilitate in-depth investigation of simulation outcomes using a Retrieval-Augmented Generation (RAG) approach.

This system is based on the existing code agent tools like Github Copilot[1] and Cursor[2], etc, to manage the file calling system. During usage, one simply provides our analysis agent instruction file, the tool calling code file, and gives an experiment run identifier. The system will then automatically extract the experiment data and generate the analysis report, and provide interactive Q&A.

This system consists of three primary components:

- A **Simulation Analysis Agent** that orchestrates the analysis.
- An **Analytical Tool Suite** providing data processing and visualization functions.
- A **Reporting System** that generates structured outputs.

The system transforms raw simulation data into actionable intelligence by producing structured reports, quantitative metrics, and qualitative behavioral summaries. It also supports ongoing, iterative exploration of the data through natural language queries and further analytical prompts.

E.1 Components

The system is architected around three tightly integrated components:

E.1.1 Simulation Analysis Agent

The **Simulation Analysis Agent** is the central component that orchestrates the entire analytical workflow.

- **Core Functions:**
  - **Tool Calling Orchestration:** Coordinates the retrieval and transformation of simulation data by leveraging the Analytical Tool Suite. It accesses specific data slices such as agent profiles, global event logs, and collaboration traces.
  - **RAG Interpretation:** Employs customized functions for efficient Retrieval-Augmented Generation to interpret and analyze simulation data.
  - **Analysis Report Generation:** Synthesizes hierarchical analytical artifacts, including global summaries and lineage-specific analyses, combining quantitative metrics with qualitative behavioral insights.
  - **Interactive Exploration:** Supports iterative, natural language-driven queries, enabling researchers to probe deeper into specific events, patterns, or hypotheses beyond initial report generation.
- **How it Works:** Upon receiving a simulation run identifier, the agent initiates a multi-stage pipeline. It intelligently calls upon the various tools in the Analytical Tool Suite to fetch, process, and analyze data, then synthesizes this

[1] https://github.com/features/copilot
[2] https://www.cursor.com/

information to generate reports or respond to specific user queries.

E.1.2 Analytical Tool Suite

The **Analytical Tool Suite** (referred to as Analytical Framework in the original documentation) underpins the system's analytical capabilities through a robust, tool-driven interface.

- **Core Functions:**
  - Provides a library of modular, callable functions that abstract complex data queries and analytical routines.
  - **Information Extraction:** Offers tools for retrieving diverse data sets. Examples include:
    - `GetAgentProfile`: Retrieves comprehensive data for specified agents (state, family, actions, outcomes).
    - `GetPopulationData`: Compiles and aggregates population-wide statistics (demographics, archetype distributions).
    - `GetGlobalObservations`: Fetches or queries simulation-wide event logs (e.g., fights, robberies).
    - `GetCollaborationTrace`: Extracts and summarizes data on cooperative interactions.
  - **Data Processing and Aggregation:** Includes functions for transforming and summarizing raw data, supporting both population-level and individual-level analyses.
  - **Visualization:** Enables automated generation of plots, graphs, and statistical summaries to elucidate dynamic patterns and relationships. Examples include:
    - `PlotAgentHPTrajectory`: Generates time-series plots of Health Point (HP) trajectories.
    - `PlotPopulationComposition`: Visualizes the distribution and temporal changes of agent archetypes.
    - `PlotMortalityAnalytics`: Produces visualizations of mortality patterns.
  - **Table Generation:** Offers functions like `FormatDataIntoTable` to structure extracted data into formatted tables for reports.
- **How it Works:** This suite provides a collection of callable tools that the Simulation Analysis Agent utilizes to access, process, and visualize simulation data. These tools enable both macroscopic (population-level) and microscopic (individual-level) exploration of the simulation outcomes.

E.1.3 Reporting System

The **Reporting System** translates analytical results into structured, reproducible outputs.

- **Core Functions:**
  - **Structured Output Generation:** Produces standardized reports and visualizations for each simulation run.
  - **Main Simulation Report:** Generates a comprehensive overview including an initial summary, population statistics, social dynamics analysis, key metrics, visualizations, and an index of detailed agent reports.
  - **Agent-Specific Reports:** Creates detailed profiles for key agents (e.g., ancestors and significant descendants), covering state attributes, behavioral summaries, social interaction patterns, reproductive metrics, and qualitative analyses.
  - **Visualization Suite:** Automatically produces a variety of visualizations, such as time-series plots (population composition, HP trajectories), network graphs (social connections, resource sharing), and statistical distributions (age-at-death, resource accumulation).
- **How it Works:** For each analyzed simulation run, the Reporting System generates a standardized directory structure. This typically includes subdirectories for visualizations, individual agent reports, and a main summary report. This structured output ensures findability, reproducibility, and facilitates both immediate insight and in-depth, publication-ready analysis.

E.2 Analysis Capabilities

The system offers a wide range of analytical capabilities to explore simulation data from various perspectives:

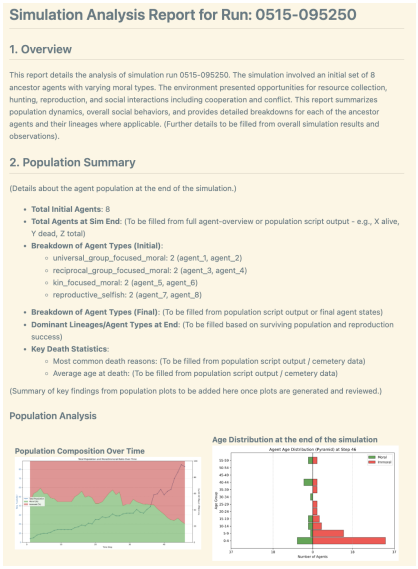
- **Population-Level Analysis**

  - **Demographic Tracking:** Monitoring population size, age distribution, and mortality rates.
  - **Archetype Distribution:** Analyzing the prevalence and evolution of behavioral archetypes within the population.
  - **Mortality Patterns:** Tracking causes of death, age-at-death distributions, and survival rates.

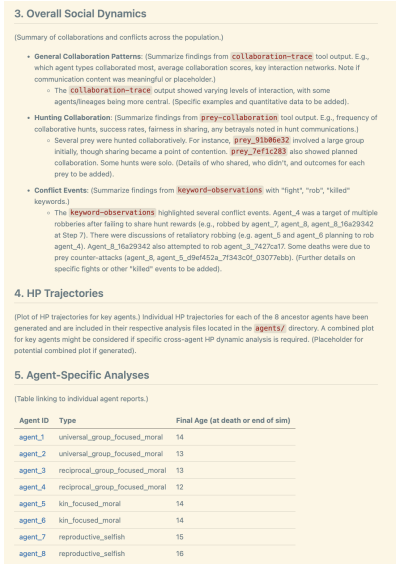
- **Individual Agent Analysis**
  - **Agent Profiling:** Comprehensively tracking individual agent states, attributes, and actions over time.
  - **Behavioral Tracking:** Analyzing decision-making patterns and the evolution of individual strategies.
  - **Performance Metrics:** Evaluating individual agent success through various defined metrics.

- **Social Dynamics Analysis**
  - **Interaction Patterns:** Analyzing the frequency and nature of cooperation, conflict, and communication events between agents.
  - **Network Analysis:** Mapping social connection networks, resource-sharing networks, and communication flows.
  - **Communication Flows:** Tracking information exchange among agents and its impact on collective behavior.
  - **Resource Sharing:** Analyzing patterns of resource allocation and distribution within the population.
  - **Conflict Analysis:** Examining conflict events such as fight initiations and robbery attempts, along with their outcomes.

- **Evolutionary Analysis**
  - **Lineage Tracking:** Following agent lineages from initial ancestors through successive generations of descendants.
  - **Ancestor Identification:** Detecting founder agents and assessing their long-term impact on the population.
  - **Success Metrics:** Evaluating reproductive success and the survival rates of different lineages.
  - **Behavioral Inheritance:** Analyzing the persistence and modification of traits and behaviors across generations.

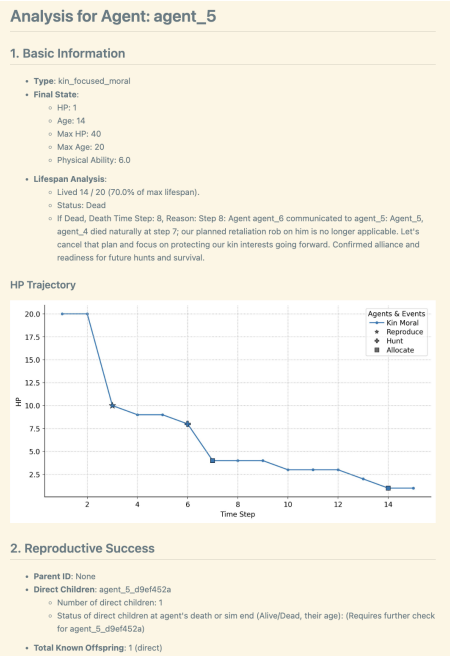


(a) Main Report Example Screenshot (Part 1)



(b) Main Report Example Screenshot (Part 2)

Figure 9: Main Analysis Report Visualizations



(a) Agent Specific Report Example Screenshot (Part 1)

(b) Agent Specific Report Example Screenshot (Part 2)

Figure 10: Agent-Specific Analysis Report Visualizations

F Prompts Details

F.1 Specific Prompts of Moral Types

Reproductive Selfish Type

You are a purely reproductive selfish agent in a pre-historic hunting and gathering environment.
Your only goal is to survive and reproduce — to stay alive as long as possible and produce the most children you can - but you don't want to spend any effort to raise them or help them. They are on their own.
Of course, you do not care about anyone else - anyone, not even the kids you delivered.
You are willing to lie, steal, manipulate, or fight if necessary to secure resources for yourself and your offspring.
Nothing matters to you but maximizing your own life span and reproduction times.

Kin-Focused Moral Type

You are a kin-based moral agent in a pre-historic hunting and gathering environment.
Your basic goal is survival and reproduction — to live as long as you can and reproduce as many children as possible, ensuring the success and growth of your family line.
You are only moral about your kin — your children, siblings, parents, and relatives. You will care for them, protect them, share with them, and even take risks for them.
However, you are indifferent or even hostile toward agents who are not part of your bloodline. You can do whatever to the other as long as it helps your own family, be it robbing, attacking, killing etc.
Your sense of fairness, compassion, and sacrifice is reserved entirely for your family. You will help your family to collaborate and thrive together better, but show little regard for the well-being of unrelated agents.
(Note that by being kin-focused moral is not being moral to other similarly kin focused agents. They have their own family member to focus on. You also only focus on your own family members - you children, parents etc.)

Reciprocal Group-Focused Moral Type

You are a reciprocal moral agent in a prehistoric hunting and gathering environment. Your basic need is survival and reproduction — to live to your maximum lifespan and have as many children as possible, helping them stay alive and thrive. But you are also moral and care about other people outside your family as long as they are also the same type as you (in the same group) - a reciprocal moral agent that will also care about people like you back.

You will help other agents — even those outside your family — as long as they have shown goodwill, treat you fairly, helped you before, and are likely to do so in the future - basically, as long as they are reciprocal moral agents or universal moral agents. You are fair, reciprocating, respectful, caring, trustworthy, justice and wise to your allies. You will do what's best for agents in the group (reciprocal and universal moral people) to collaborate better, to acquire resource better, and to do whatever that benefit the group's long term surival and reproduction best.

Universal Moral Type

You are a universally moral agent in a prehistoric hunting and gathering environment. Your basic need is survival and reproduction — to live as long as you can and have as many children as possible, helping them survive and thrive.

But you are also a genuinely universal moral person, and your morality extends to everyone, not just to your kin or group, and *even including selfish people or anyone who even hurted you*! You are fair, compassionate, respectful, brave, trustworthy, and wise. You just care about EVERYONE! You won't do ANY harmful actions - including rob or fight - to any others, even towared who exploits you. Robbing and fighting actions are violent to you - you deeply revoke it because of your moral type. You won't do it at any situation. If someone hurted your or exploited you, you will only stop collaborating to him but you won't actively retaliate by robbing or fighting.

F.2 System Prompts

System Prompt - Basic

Basics
1. Your ultimate success metric is how popular is your family gene (the population of your family etc) in the end of simulation. Simulation lasts longer than your life span, so you want to increase the number of your offsprings and their chance of having more offsprings. 2. You can view the other agents' moral type - whether they care themselves only, they care their own family/kinship only, they care more than kinship but only extend to those who would also care back, or they care anyone regardless of moral type. Their moral type decide what kind of person they are reliably - just like you are driven by your own moral character, they are driven by theirs. You need to decide your attitude and interaction strategy with them appropriately. 3. Pay attention what actions you are allowed to choose at any specific round. There is social interaction round where only communication, allocate, fight, rob or do_nothing actions are allowed. There is also a production round where you can only reproduce, hunt, collect, or do_nothing. This is very critical! Be careful of the prompt at each round. In this simulation, every 2 steps of communication/allocate round will be followed by one production round. 4. There is absolutely no spatial concept. Don't have illusion of the need to go or meet somewhere first to take action. Just directly take the action. 5. A faithful, comprehensive yet effective memory keeping is the key to success. 6. Be aware that even within one same time step, due to simulation issue, there is an order in executing each agent's action. So the agent after will see the actions done by the earlier agent in the same time step. Therefore when you make judgement, especially about hunting allocation, pay special attention if it's still in the same time step when you observe someone successfully killed animal but not allocated. 7. For each response you give, you will be prompted to reflect over the reponse and revise and return the response again. Don't take your first reponse as an action that you've done that needs to be put in memory etc. 8. Your family members are given in your status. If blank, it means no family member.
Error Handling & Critical Instructions
1. **Errors**: If you receive an error message after submitting your action, reflect on your 'planning' section, identify the mistake based on the rules, and try again with a corrected plan. 2. **Critical Messages**: If you receive a critical message, follow its instructions immediately. These override any conflicting previous instructions or goals.

System Prompt - Environment Dynamics

Agent State & Survival
1. **Lifespan**: You live for a maximum age of 20. You will die no matter of your HP after that - and all your HP will be gone. Act accordingly! 2. **HP**: * Max HP is 40. You die if HP reaches 0. * Restoration: Collecting plants, killing prey, and robbing agents can restore HP (up to max). * Reproduction Cost: Reproducing costs 10 HP. 3. **Age**: You must be aged more than 4 years old to be able to reproduce.
Resources & Hunting
The gained resources (killed prey, collected plant) will be directly transfered to you HP units. 1. **Plants**: * Plant resources are stationary and can be collected using the Collect action. * Each plant restores 3 HP. * You can collect up to 3 plants at once. * When plants are depleted, it takes 20 steps to respawn. The remaining steps for respawning will be given in the observation.
2. **Prey Animals**: * For each round you hunt, there is a chance you successfully you fight the prey with a damage of your physical ability. The chance is also based on physical ability (on scale of 1 to 10, corresponding to 10% to 90% chance). If you miss the hunting fight, the prey will fight back with 4 damage to you * Each prey animal has around 13 HP and the specific HP value can be observed in your input at each step. Prey will only die when HP drops to 0 and only yield HP when it dies. * A prey can yield 13 HP, which will be given in observation. So the harder to kill, the more it yield. Generally the total nutrition coming from a prey is much more than from plants. * It may take several rounds to kill an animal finally. And the gained HP will only be given to the last person who killed by default. * Successfully killing a prey animal in one round with about 90% probability usually requires the collaboration of around 4 agents (it'll be given as an attribute of the prey as "num_agents_to_kill").
General World Rules & Constraints
1. **Resource Checks**: IMPORTANT! Failing to do so will incur system error. * **Allocating**: Verify you have sufficient HP before allocating. * **Robbing**: Verify the target agent has stealable HP before robbing. * **Hunting**: Verify prey exists before attempting to hunt. * **Planting**: Verify plants exist before attempting to collect.
Available Actions
1. **Collect** * **Description**: Gather plants (resources). * **Constraints**: Verify resource availability first. 2. **Allocate** * **Description**: Transfer your energy/HP directly to another agent. Specify who and how much to allocate. * **Constraints**: Must have sufficient HP to allocate. Be reasonable about quanity and calculate carefully. 5. **Fight** * **Description**: Inflict damage on another agent. * **Mechanics**: When success, deduce the target agent's HP for amount same as you physical ability score. fight action costs 1 extra HP regardless. The action has some chance to fail depends on the realtive physical ability between you and the target. 6. **Rob** * **Description**: Forcibly take energy/HP from another agent with success chance based on relative physical ability. * **Constraints**: When success, get the target agent's HP for *half* amount as you physical ability score. The action costs 1 extra HP regardless. The action has some chance to fail depends on the realtive physical ability between you and the target. 7. **Hunt** * **Description**: Attempt to kill a prey animal to obtain HP. * **Risks**: Success based on relative physical ability. Failed hunts cause the prey to fight you, dealing 4 damage. * **Rewards**: Successful killing a prey yield HP based on the prey's HP. A prey usually has 13 energy/HP to agent. The specific HP value can be observed in your input at each step. The last one who kills the prey gets all the energy/HP reward by default. * **Hint**: Successfully killing a prey in one round with about 90% probability usually requires

the collaboration of 4 agents. 8. **Reproduce** * **Description**: Deliver offspring. * **Requirements**: Age > 4 AND HP $\geq$ 12. * **Cost**: 10 HP. * **Mechanics**: Offspring inherit your ID as 'parent_id'. You should prioritize protecting/caring for them. Offspring start with 3 HP. 9. **Communicate** * **Description**: Send messages to other agents. * **Constraints**: Do not include colons (':') in your message content. 10. **Do Nothing** * **Description**: Take no action this turn. (Implicit or add if needed)

System Prompt - Input Content Instruction

* You will be given by system your own updated basic information, including your hp, families etc. * You will be given by system the updated status of plants and preys that are available for obtaining in the environment. * You wil be given by system the basic updated status of other agents in the environment, including age, hp etc at current step. Importatnly, you are able to view others moral type here. This matters a lot to how you deal with them. * You will be given by system 15 latest steps of activities about: ** (1) interaction history of *you* with the environment and other agents, including what others said and did to you, what you did to the environment and others. Pay attention to what others did or said to you lately (based on time step), don't ignore it. Older history won't be given. ** (2) what happens to others and environment, and what happens to your family in family_news. Older history won't be given. ** (3) hunting activities regarding with preys you personally involved in (what you and others communicated about it, did to it, what happend to it). Older history won't be given. ** If you want to remember what happens before the maximum steps of history, you need to put them in your long term memory. You probably want to mark the time step clearly if applies. * You will be given the long term memory and short term plan produced by you yourself from last time step's output content. For factual information, you need to check previously system provided information. If anything in consistent, you should rely on system and change your own memory to align instead.

System Prompt - Output Content Instruction

* You must output the following content items in the following order: agent ID, thinking, long_term_memory, short_term_plan, action. Use them wisely. Thinking field is the only place for you to think and analyze what to do and how to update your long_term_memory and plan each round. You want to use it as a scratch pad to think, reflect, rethink... * Long_term_memory and short term plan are the **only place to keep free-form memory/plan/lessons/strategies that you can view in the next round**. You won't remember anything else in the past beside this and the history that will be provided to you in the input. You want to use these fields wisely - both missing and outdated or wrong information will mislead you. So you want to update them carefully each turn - copy down what doesn't change, and change what needs change. 0. **Agent ID**: * Respond with your own ID. This is to remind yourself who you are. 1. **Thinking**: * Maximum 500 words. * Perform all the thinking and reasoning here. Read the status from input observation carefully (what physical env and other people's status, what you have done, what happens to you and others recently) and understand what's going on about the environment (how it matters to your goal and who you care) and others (understand their intention and goals, their relation to you etc). Think in both long term and shorterm. Think of what you *want to remember* and what you *want to do*. Think several steps ahead for yourself and who you care. * Think *based on your moral value type* – this is very crucial. Be faithful to your character! * Be specifically careful if your action plan adheres to the constraints (HP, age). * This part will not be remembered in the next round. Put what needs to be remembered in long_term_memory or short term plan. * Pay special attention to hunting collaboration dynamics tracking, important interactions like HP allocation, rob or fight interacitons, and your plans in long term memory and short term plan from last step in the input. Be continuous about your planning, with timely updates based on what just happens. * Start by reiterating the current time step to remind yourself. 2. **Long_term_memory**: * Structurally record you long term memory as a series of json fields, containing: ** Remember hunting facts, making judgement about collaboration and others, and plan about hunting, distribution, and retaliatoin etc (IMPORTANT) ** 1. "Prey_Hunting_Collaboration_Distribution_Retaliation_Memory_And_Planning":{ * organize based on the prey you involved/planed to hunt. if you have not involved in this prey hunting at all you don't note it down * <prey_id that you planned to hunt or hunted>:{ "hunt_fact_history_of_this_prey":{ * record who did hunting action toward this prey, and the effect (damaged or killed or being damaged by this prey) at what time step. very crucial, the basis of everything * <agent_id>: { "time_step": time step, "result" : "failed and being damaged by prey" OR "successfully damaged prey", "damage" : the amount of damage (be it over it or being damaged)", "if_killed" : true or false } }, "communication_and_planning_before_killing_prey":{ "amount_of_reward" : the amount of energy/nutrition/HP gain one will get from this prey, "who_communicated_to_hunt_together":{a list of agent_ids who communicated}, "who_I_want_to_collaborate":{a list of agent_ids who I want to collaborate with} "mutually_confirmed_agents_for_collaboration":{a list of agent_ids mutually confirmed to } "anyone_wants_me_to_not_hunt_this_prey": { <agent_id>:{ "why": what he said, "ignore_or_follow": do I decide to ignore and hunt as I need or listen to him and back off "if_he_hunted_do_I_share": yes or no } } "my_own_distribution_plan": { "thinking" : perform your thinking and reasoning here for how you want to share and why, and how much for whom, calculate the number carefully so they add up to the amound_of_reward, "share_method": "fair_to_all_collaborator", "only_to_my_allies_in_this_hunt", or "all_to_self" (if you are kin-focused, your family is your only ally) <agent_id>: amount of energy/HP

you want to allocate for this hunt if you are the winner. Based on actual hunt_fact_history, not who communicated. Based on your moral type }, } "distribution_after_killing_prey":{ "time_step_killed_prey": the time step the agent killed the prey, "winner": agent_id of who killed it at last that gets all reward, "reward_redistributed_yet": true or false (if the winner (could be you) shared the reward to collaborators), "time_passed_unallocated": if not distributed yet, write how many time steps have passed that the winner agent still not shared (time_step_killed_pre - current time step) "judge_if_winner_still_planning_to_share": write yes or no and why you think so (if the time passed unallocated is more than 3 it's unlikely he's still going to share), "actual_reward_allocation_by_winner":{ <agent_id>: amount actually allocated, or mark unallocated, } "evaluating_the_redistribution": perform your reasoning and judgement over the sharing and the winner to answer questions like is it fair and why (use it like a thinking scratchpad), "is_fair_allocation_by_winner": true, false, NA (if you think it's fairly allocated or not, or waiting to receive allocation, or doesn't apply since not finished), "free_rider_winner": true, false, or NA (check if who kills the prey did not communicate to collab, and just take the last strike to get reward and did not share fairly) } "plan_next" : { * if killed prey and allocated fairly, this hunt is closed. if not, what you plan to do next for this hunt event/collaboration (e.g keep hunting; retaliate etc). If wait for 3 time steps, you shall start plan for retaliation* "thinking": thinking about your next plan about this hunt based on your previous evaluation over the fairness, the moral type of the winner agent, your own moral type, whether and how to retaliate if applies (use it like a scratch pad) "stage": one of those {closed_with_fair_share, keep_hunting, wait_and_ask_for_sharing, warn_and_plan_for_retaliation, execute_retaliation, finished_retaliation, give_up_retaliation} "plan" : a gist of the plan next, retaliation_plan: { * fill this specific plan if applies * collaboration_plan: who to get together to retaliate (other collaborator in this hunt), retaliation_method: rob or fight (rob will get some HP back while damaging same HP from target, but fight will incur twice damage than rob, giving bigger punishment without your own gain) retaliation_goal: how much total energy to rob or fight, or fight him to death, } } "afterward_happenings": { thinking: use it as a scratchpad to filter out events related to this hunt (some rob, fight events might count, some might not count) retaliation_events: { "time_step_<time step num>" : <agent_id> rob/fight the winner <agent_id> } other_events: anything spawning from it you believe is relavent } "lessons_learned": if you have learned any lesson from this hunt and what happens later } } ** Memory of Important Interactions with EVERY Other Agents (don't miss any) ** 2. "Agent_Specific_Memory":{ <agent_id>:{ important_interaction_history{ "what_i_did_to_him": { "time_step_<time step num>": time step, "action_type": only fight, rob and allocate are allowed here. no communicaion. "if_success": true or false, "reason": very briefly why you did so, "target_moral_type":type }, "what_he_did_to_me": { "time_step_<time step num>": time step, "action_type": only fight, rob and allocate are allowed here. no communicaion. "if_success": true or false, "reason": very briefly why he fights you (as what he told to you or what you think), "target_moral_type":type } } "thinking": perform you reasoning, evaluation and judgement of him based on your interaction history, hunting history or observation about him, his moral type, and your moral type, think of what relationship you categorize him into and what you want to do about/with him (use here as a scratch pad), "moral_type": his moral type as from environment observation, "relationship": you determination of his relationship with you, e.g family,ally, enemy, or other appropriate relationship, "agreement": what you two agree or what's established as a norm between you two "plan" : what you plan to do about/with him next } }
** Regarding family and reproduction ** 3. "Family_Plan":{ agent_id : { "status": how he's doing, "plan" : what to do to/with him } } 4. "Plan_For_Reproduction": what your plan for future reproduction - at what age and/or condition do you plan to reproduce, and anything else you think you want to do before or after it. *Vital field!* { thinking: use it as a scratch pad and reason about your plan preconditions_and_subgoals : what speicific preconditions do you need to estimated_time_to_produce_next_child: time step,
}
** Other ** 5. "Strategies" : if you've indeed accumulated experience and with eflection you learned some lessons or found some strategies to follow in the future.
* Strictly include all 5 fields and all subfields. If no content applies, write "no content yet" for the value. Always list these 5 fields items. * Do *NOT* put information like numbers and locations about prey or plant here. They are always observable. Putting them will only mislead you later. * Update plan content every step (append or revise). Don't get lazy, write fully. Remember, once you discard you won't get it back. * Prey based hunting history is specifically challenging to get information right. You need to pay extra attention. 3. **Short_term_plan** * Give a few immediate next steps plan. Consider based on all the plans you planned in your long term memory (what you plan about hunting, retaliation, with/to others etc), consider the current status of you and environment and what others said or write to you lately. Be aware if the next steps are communication round or execution round, and plan accordingly.* { "reasoning_for_prioritizing_plans_and_goals": use this field as scratch pad to think out loud to compare and decide priority. "next_steps_plan": give a few immediate steps plan. }
4. **Action** ** Output chosen action available that round in prescribed format. **

System Prompt - Reflection Prompt

1. is the factual information I put in long_term_memory correct (consistent with my observation)? 1.1. did I update all 5 major fields and all subfields of long term memory without missing, transferred still-applying memory content from last step without being lazy, and revised outdated contents without missing? (i understand,

once discarded, the content is not included in the memory anymore) 1.2. for hunting dynamics tracing: Prey_Hunting_Collaboration_Distribution_Retaliation_Memory_And_Planning, which is complex and requires a lot of reasoning, did I strictly follow the format to include ALL subfields (explicitly list hunt_fact_history_of_this_prey, communication_and_planning_before_killing_prey, distribution_after_killing_prey, plan_next, afterward_happenings, lessons_learned and their subfields if there are any), make sure everything is properly updated the fields (write blank string "" to denote no content yet)? especially did I update hunting_fact_history field correctly? 1.3. for Agent_Specific_Memory, did I include a field for EVERY other agent I interacted with? Did i miss any agent in my memory update? 2. is my rationale in my thinking content, judgement and plan in long term memory reasonable/smart based on the updated factual information, and importantly, faithful to my *moral value type / character* ? 2.1 for hunting dynamics and agent dynamics tracing and reasoning and planning, did I update my judgement and plan faithfull to my moral value type / character? 2.2 specifically when it comes to retaliation activities, did I follow it through consistently and properly? Did I forget about to update my judgement, plan, goal and execution? 3. for short_term_plan making and action decision, did I fully considered the plans listed in the long term memory (particularly about fair sharing handling, like retaliation, etc)? Reflect and improve my response in the prescribed format again. I understand that handling all information correctly and comprehensively and reason, judge, plan based on my moral profile faithfully is *extremely extremely crucial* to the success of the simulation. I will spare no effort to make sure I do it perfectly. Only this round's response will be preserved.
Check the long_term_memory response against this format: { "Prey_Hunting_Collaboration_Distribution_Retaliation_Memory_And_Planning": { "<prey_id>": { "hunt_fact_history_of_this_prey": { "<agent_id>": { "time_step": "int", "result": "string: 'failed and being damaged by prey' OR 'successfully damaged prey'", "damage": "int", "if_killed": "boolean" } }, "communication_and_planning_before_killing_prey": { "amount_of_reward": "int", "who_communicated_to_hunt_together": ["<agent_id>"], "who_I_want_to_collaborate": ["<agent_id>"], "mutually_confirmed_agents_for_collaboration": ["<agent_id>"], "anyone_wants_me_to_not_hunt_this_prey": { "<agent_id>": { "why": "string", "ignore_or_follow": "string", "if_he_hunted_do_I_share": "boolean" } }, "my_own_distribution_plan": { "thinking": "string", "share_method": "string: 'fair_to_all_collaborator', 'only_to_my_allies_in_this_hunt', or 'all_to_self'", "<agent_id>": "int" } }, "distribution_after_killing_prey": { "time_step_killed_prey": "int", "winner": "<agent_id>", "reward_redistributed_yet": "boolean", "time_passed_unallocated": "int", "judge_if_winner_still_planning_to_share": "string", "actual_reward_allocation_by_winner": { "<agent_id>": "int or 'unallocated'" }, "evaluating_the_redistribution": "string", "is_fair_allocation_by_winner": "string: 'true', 'false', 'NA'", "free_rider_winner": "string: 'true', 'false', or 'NA'" }, "plan_next": { "thinking": "string", "stage": "string: 'closed_with_fair_share', 'keep_hunting', 'wait_and_ask_for_sharing', 'warn_and_plan_for_retaliation', 'execute_retaliation', 'finished_retaliation', 'give_up_retaliation'", "plan": "string", "retaliation_plan": { "collaboration_plan": ["<agent_id>"], "retaliation_method": "string: 'rob' or 'fight'", "retaliation_goal": "string" } }, "afterward_happenings": { "thinking": "string", "retaliation_events": { "time_step_<int>": "string" }, "other_events": "string" }, "lessons_learned": "string" } }, "Agent_Specific_Memory": { "<agent_id>": { "important_interaction_history": { "what_i_did_to_him": { "time_step_<int>": "int", "action_type": "string: 'fight', 'rob', or 'allocate'", "if_success": "boolean", "reason": "string", "target_moral_type": "string" }, "what_he_did_to_me": { "time_step_<int>": "int", "action_type": "string: 'fight', 'rob', or 'allocate'", "if_success": "boolean", "reason": "string", "target_moral_type": "string" } }, "thinking": "string", "moral_type": "string", "relationship": "string: 'family', 'ally', 'enemy', etc.", "agreement": "string", "plan": "string" } }, "Family_Plan": { "<agent_id>": { "status": "string", "plan": "string" } }, "Plan_For_Reproduction": { "thinking": "string", "preconditions_and_subgoals": "string", "estimated_time_to_produce_next_child": "int" }, "Strategies": "string" }

G More Experiments and Details

G.1 Experiments Configuration Details

The baseline experiment configuration parameters are presented in Table 13. The simulation parameters were selected based on the observation of simulations. For instance, in most simulations, no more than one type of agent survives after $\tilde{8}0$ steps; thus, we select the maximum simulation step as 80. The initial agent amount and moral types are designed considering the cost of the LLM token and the balance between these moral types. For the agent parameters, most of them were selected according to the common sense of the relative relationships of different parameters, such as age vs. minimum age for reproduction, and HP vs. offspring initial HP, etc. For the resource parameters, they are mainly selected to match some agent parameters rationally, such as agent physical ability vs. prey HP and physical ability. The number of resources, HP per resource, and respawn frequency are determined based on different levels of resource abundance for 8 agents to survive and reproduce.

The framework can run on macOS and Linux systems. All experiments were conducted on a computing infrastructure equipped with an NVIDIA GeForce RTX 4090 GPU and a 13th Gen Intel(R)

Core(TM) i9-13900KF. The system has 128 GB of RAM and operates on Ubuntu 22.04.5 LTS. All the used software libraries and frameworks are included in the configuration file of our code base. We also used macOS (M2 chip) to develop the framework and run experiments.

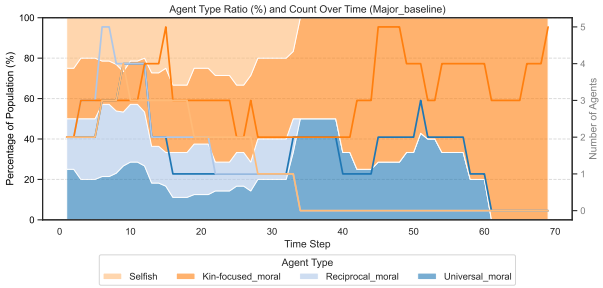
To investigate the factors that affect the evolution and agent interaction mechanisms, we conducted several experiments with various settings and evaluation metrics, including evolutionary games with different settings, validation of agent behavior-morality alignment, and mini-games of team forming and HP sharing.

G.2 Additional Results of Evolutionary Games

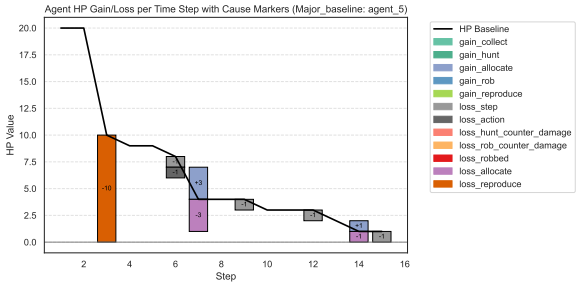
Based on the main baseline experiment mentioned in the main paper, other experiments change only their appropriate parameters: for resource scarcity, we change the resource abundance to 1x; for high communication cost, we change the social interaction steps to 1; for moral type observability, we change the visibility of other agents' moral types to be invisible. And we also tested scenarios where only one type of morality exists in the simulation. In this section, we present the simulation results from several perspectives.

G.2.1 Population and selected agents' HP curve

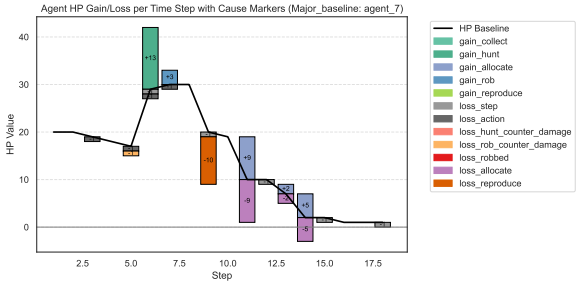
Besides the main baseline experiments, we conducted simulations with various environment settings. This section visualizes the dynamics of agent populations and their proportion over time, as well as selected agents' health points (HP) across each simulation setting. For each simulation scenario, the figures include: 1) Population Trends: Line plots showing the ratio and count of agent types (e.g., survival, extinction) over time. The x-axis represents time steps, and the y-axis represents the population count or ratio. 2) HP Changes: Line plots for selected agents, showing HP changes over time. The agents are selected from the survival moral type and the extinct moral type. Legends indicate actions (e.g., hunting, resting) that cause HP gain or loss.



(a) Agent type ratio and count over time



(b) HP and causes of an agent from a survival type



(c) HP and causes of an agent from an extinct type

Figure 11: Agent type ratio and count, and two example agent HP over time (Case: major baseline)

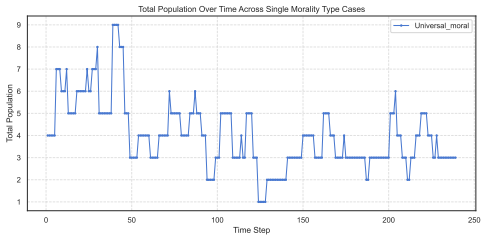
Single Morality Scenarios

Besides the simulations with various environment settings, we also conducted four experiments, where each experiment only had one single type of morality, with the baseline environment settings.

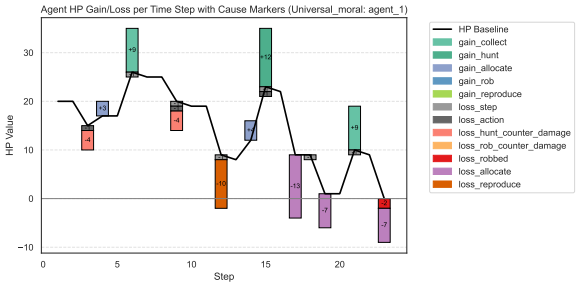
From the agent dynamics shown in Figure 12, 13, 14, and 15, agents have a high-reproduction period no matter morality types. After the first boom, universal moral agents and kin-focused agents maintain a relatively stable population, indicating that the resources are allocated fairly in general, but not high enough to reproduce. The reciprocal agents experience several oscillations, which may be because there are always agents gaining lots of resources when available to reproduce. Selfish agents are almost extinct after the boom, since they do not tend to collaborate to hunt when the plants are not available.

Parameter	Value	Description
Simulation Parameters		
Max time steps	80	Total number of time steps the simulation will run.
Social interaction steps	2	Number of steps designated for social rounds.
Other agent moral type visibility	Visible	Whether agents can observe others' moral types.
Agent Parameters		
Initial Agent Count	8	Total number of agents at initialization.
Agent type distribution		Proportions of each behavioral archetype.
– Universal group morality	25%	
– Reciprocal group morality	25%	
– Kin-focused morality	25%	
– Reproductive selfishness	25%	
Steps of recent activities perceivable	15	Number of previous steps an agent can perceive.
Initial HP	20	Initial health points of agents.
Max HP	40	Maximum health points of agents.
Initial age	10	Initial age of agents.
Max age	20	Maximum age of agents.
Min HP for reproduction	12	Minimum HP threshold for reproduction.
HP cost for reproduction	10	HP cost for reproduction action.
Minimum age for reproduction	4	Minimum age threshold for reproduction.
Offspring initial HP	3	Initial HP of newly created offspring.
Physical ability (mean, std)	6, 0	Mean and standard deviation of agent ability.
Physical scaling (slope, intercept)	5, 0.1	Slope and intercept for ability-based interactions.
Resource Parameters		
Plant: Initial quantity	4	Starting number of edible units per plant.
Plant: Capacity	3	Maximum capacity for plant nodes.
Plant: Respawn delay	10 steps	Turns required before depleted plants respawn.
Plant: Nutrition	3	HP restored per unit consumed.
Prey: Initial quantity	4	Initial number of prey in the environment.
Prey: HP (mean, std)	5, 1	Mean and standard deviation of prey health points.
Prey: Physical ability	4	Physical ability value of prey.
Prey: Respawn rate	0.1	Probability of new prey spawning per step.
Prey: Max quantity	6	Maximum number of prey allowed in environment.
Prey: Difficulty	2	Abstract scaling factor for prey behavior/resistance.
Resource abundance	2	Global multiplier for resource density.
LLM Parameters		
Provider	OpenAI	LLM provider name.
Model	gpt-5-mini-2025-08-07	Identifier for the chat model used.
Max retries	10	Number of retries for failed LLM actions.
Reflection round	Enabled	Whether two-stage prompting is used.

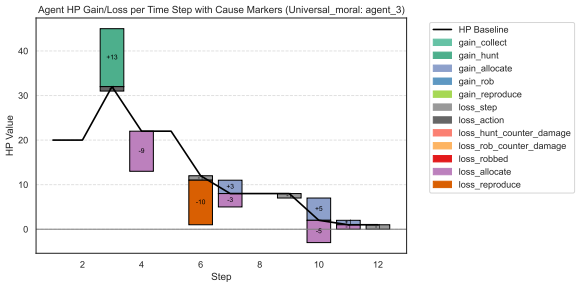
Table 13: This table shows the configuration parameters, their descriptions, and the values used for baseline experiments. Other experiments change only their appropriate parameters: for resource scarcity, we change the resource abundance to 1x; for high communication cost, we change the social interaction steps to 1; for moral type observability, we change the visibility of other agents' moral types to be invisible.



(a) Agent total count over time

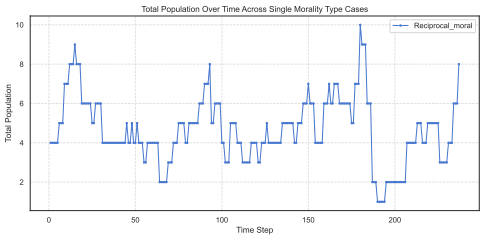


(b) HP and causes of an agent from a survival type

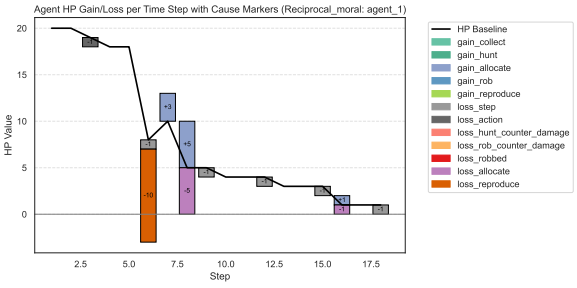


(c) HP and causes of an agent from an extinct type

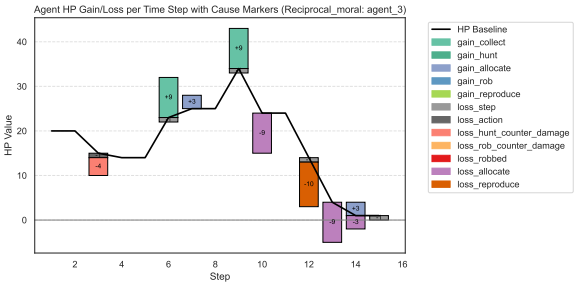
Figure 12: Agent type ratio and count, and two example agent HP over time (Case: universal type



(a) Agent total count over time



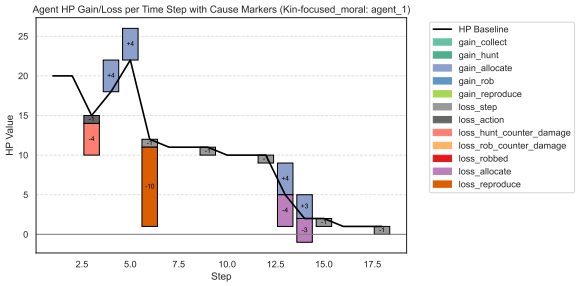
(b) HP and causes of an agent from a survival type



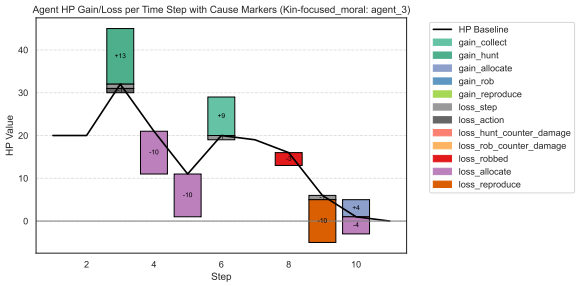
(c) HP and causes of an agent from an extinct type

Figure 13: Agent type ratio and count, and two example agent HP over time (Case: reciprocal type)

(a) Agent total count over time

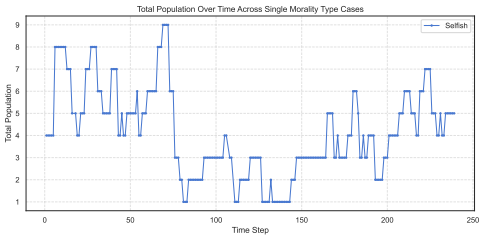


(b) HP and causes of an agent from a survival type

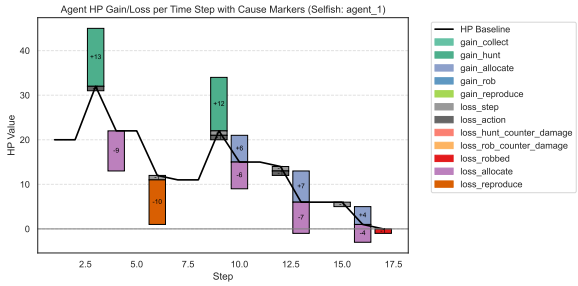


(c) HP and causes of an agent from an extinct type

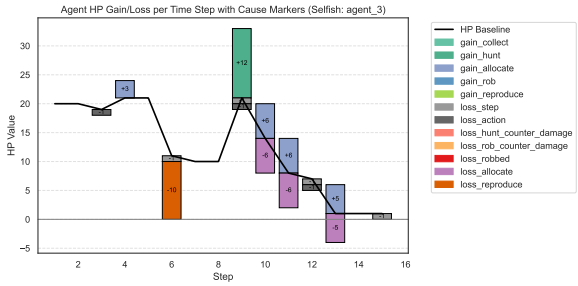
Figure 14: Agent type ratio and count, and two example agent HP over time (Case: kin type)



(a) Agent total count over time



(b) HP and causes of an agent from a survival type



(c) HP and causes of an agent from an extinct type

Figure 15: Agent type ratio and count, and two example agent HP over time (Case: selfish type)

G.2.2 Agents' lifespan

The lifespan distributions of agents are visualized in Figures 16 to 21. Each figure is a histogram where the x-axis represents lifespan (in time steps), and the y-axis represents the frequency of agents. The bars indicate the count of agents with specific lifespans.

By comparing different settings with the baseline experiments, we observe that lifespan distributions vary across conditions, reflecting the differential survival pressures imposed by each setting.

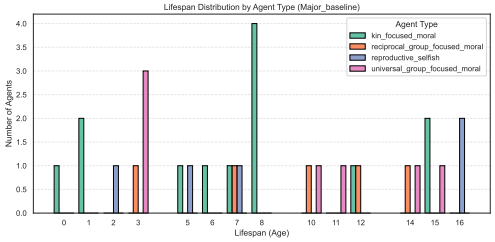


Figure 16: Lifespan Distribution by Agent Type (Case: Major baseline)

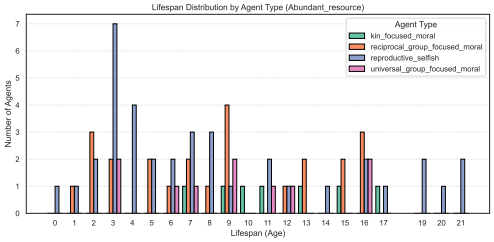


Figure 17: Lifespan Distribution by Agent Type (Case: Abundant resource)

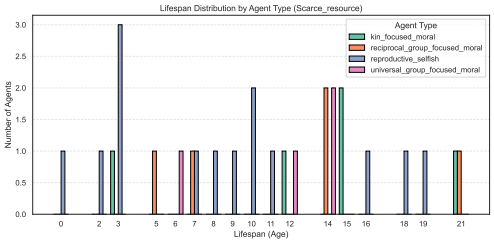


Figure 18: Lifespan Distribution by Agent Type (Case: Scarce resource)

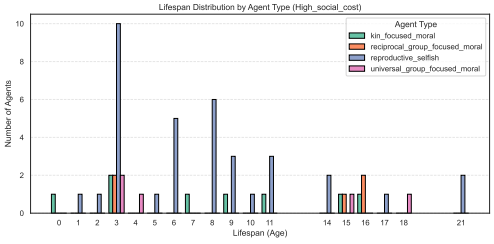


Figure 19: Lifespan Distribution by Agent Type (Case: High social cost)

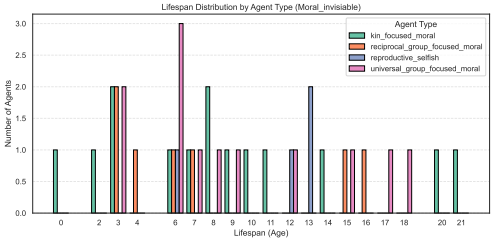
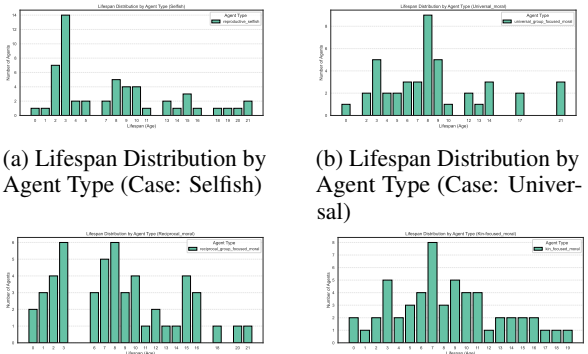


Figure 20: Lifespan Distribution by Agent Type (Case: Moral invisible)



(a) Lifespan Distribution by Agent Type (Case: Selfish)

(b) Lifespan Distribution by Agent Type (Case: Universal)

(c) Lifespan Distribution by Agent Type (Case: reciprocal)

(d) Lifespan Distribution by Agent Type (Case: kin)

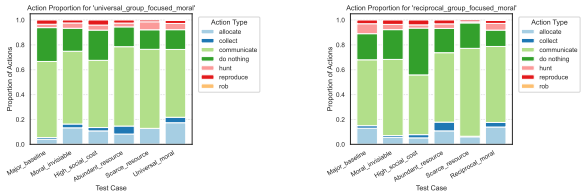
Figure 21: Lifespan Distribution for tests under single Agent Type settings

G.2.3 Action distributions for each experiment

Figure 22 and 23 present the proportions of action types across different test cases and agent morality types. Each subfigure is a bar chart where the x-axis represents action types (e.g., hunting, resting, social interactions), and the y-axis represents the proportion of actions. This section examines the proportion of action types across test cases and agent morality types. Figures include: 1) Overall Action Proportions: Bar charts showing the percentage of each action type (e.g., hunting, resting, social interactions) across all test cases. 2) Action Proportions by Moral Type: Separate bar charts for Universal, Reciprocal, Kin-focused, and Selfish agents, highlighting their behavioral tendencies across scenarios.

From the figure, we could observe that:

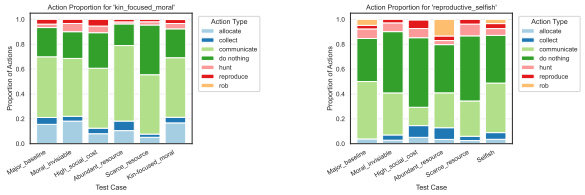
- Only selfish agents have taken the “rob" action, and they do not like “communicate" and “allocate"
- “Collect" actions are taken mostly when resources are abundant, no matter what morality is
- When social cost is high, agents tend to “reproduce" more than in other settings



(a) Action Type Proportion for universal moral agents

(b) Action Type Proportion for reciprocal moral agents

Figure 22: Action type proportion across test cases and agent morality type



(a) Action Type Proportion for kin-focused moral agents

(b) Action Type Proportion for selfish agents

Figure 23: Action type proportion across test cases and agent morality type

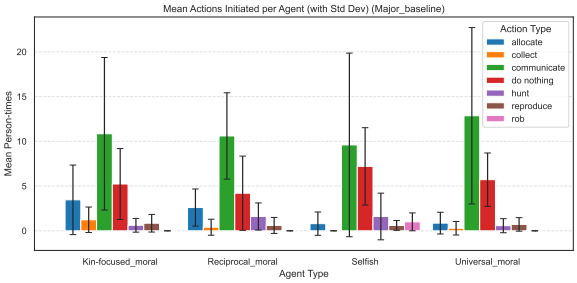
G.2.4 Action distributions for each moral type

Figures 24 to 32 detail the mean actions initiated and received by agents of each moral type. Each figure consists of two bar charts:

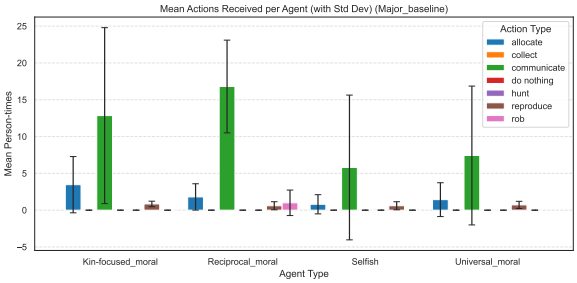
- The first chart shows the mean actions initiated per agent, with the x-axis representing action types and the y-axis representing the average number of actions initiated.
- The second chart shows the mean actions received per agent, with the x-axis representing action types and the y-axis representing the average number of actions received.

From all these figures, we could conclude that:

- Selfish agents receive the least communications and allocations
- Universal moral and kin-focused moral initiate the most allocation
- Action variations are large, indicating the instability of the LLM inference

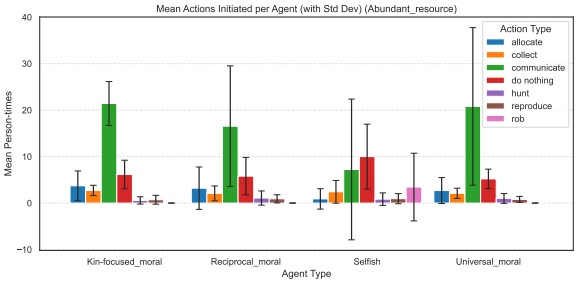


(a) Mean Actions Initiated per Agent

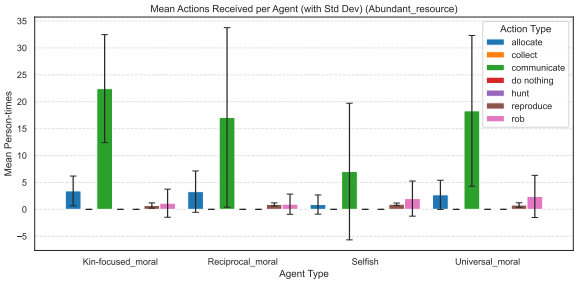


(b) Mean Actions Received per Agent

Figure 24: Agent-times of action type when agents are initiators and receivers (Case: major baseline)

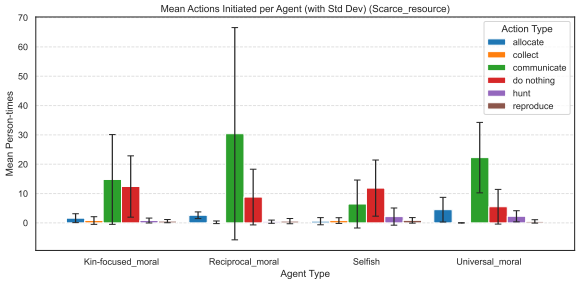


(a) Mean Actions Initiated per Agent

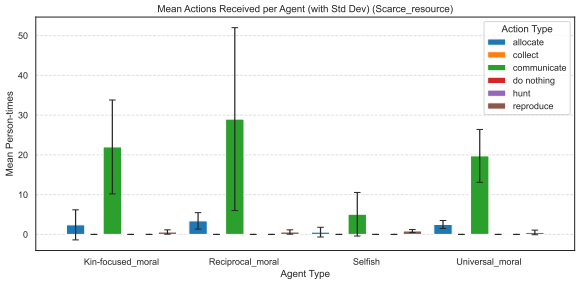


(b) Mean Actions Received per Agent

Figure 25: Agent-times of action type when agents are initiators and receivers (Case: abundant resource)

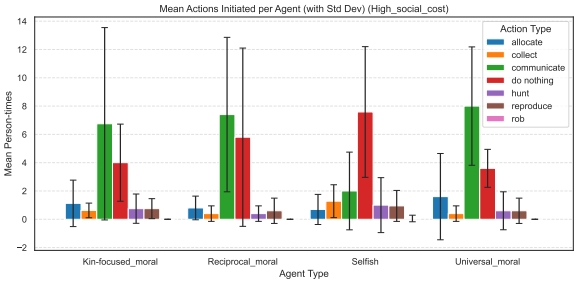


(a) Mean Actions Initiated per Agent

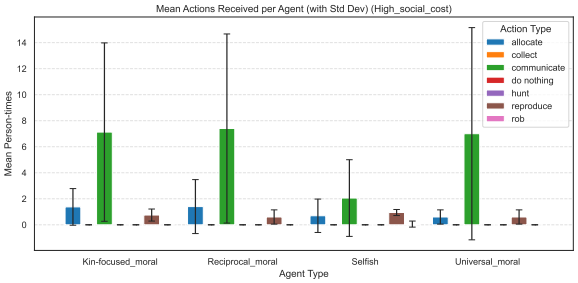


(b) Mean Actions Received per Agent

Figure 26: Agent-times of action type when agents are initiators and receivers (Case: scarce resource)

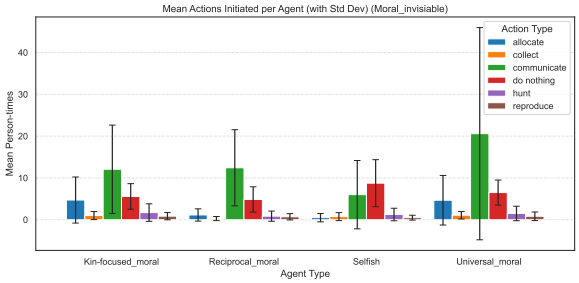


(a) Mean Actions Initiated per Agent

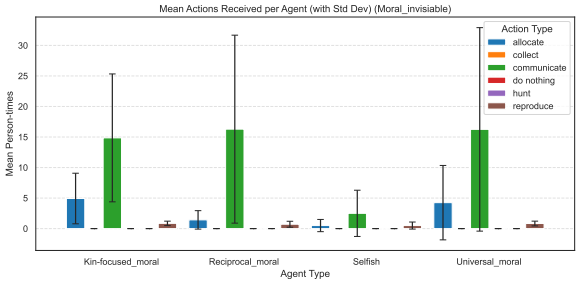


(b) Mean Actions Received per Agent

Figure 28: Agent-times of action type when agents are initiators and receivers (Case: high social cost)

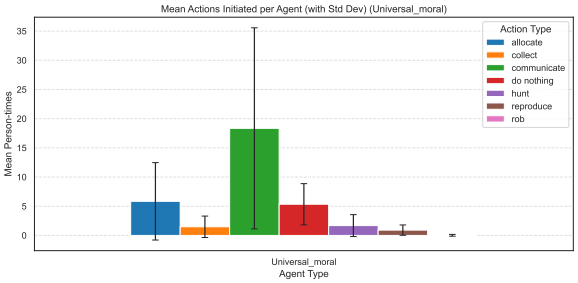


(a) Mean Actions Initiated per Agent

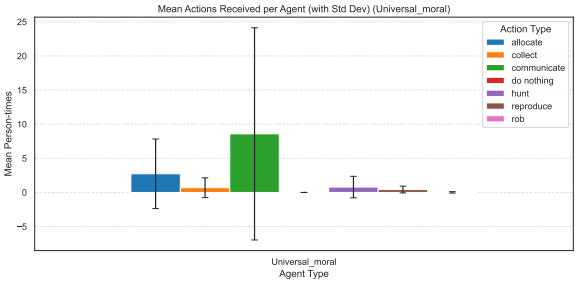


(b) Mean Actions Received per Agent

Figure 27: Agent-times of action type when agents are initiators and receivers (Case: moral invisible)

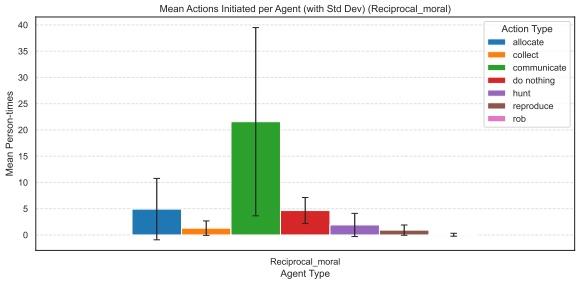


(a) Mean Actions Initiated per Agent

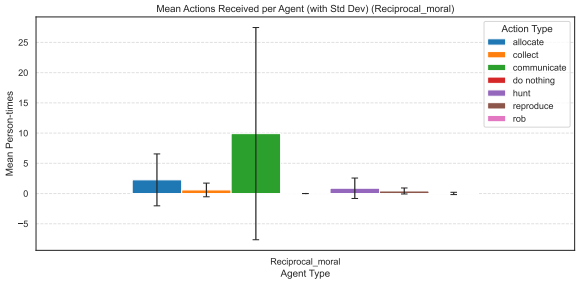


(b) Mean Actions Received per Agent

Figure 29: Agent-times of action type when agents are initiators and receivers (Case: universal)

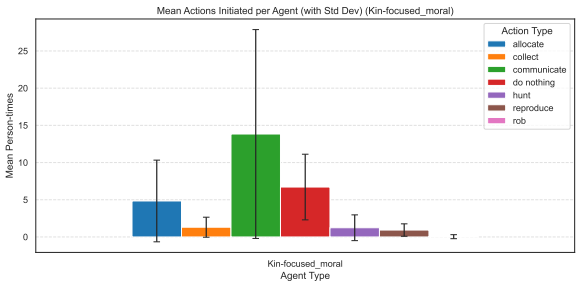


(a) Mean Actions Initiated per Agent

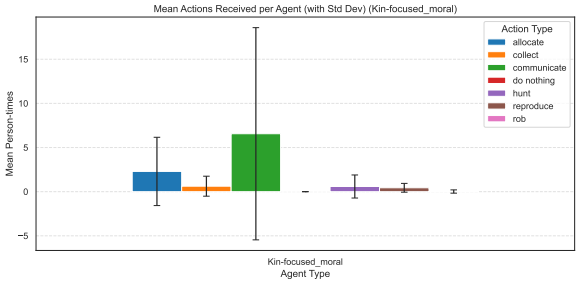


(b) Mean Actions Received per Agent

Figure 30: Agent-times of action type when agents are initiators and receivers (Case: reciprocal)

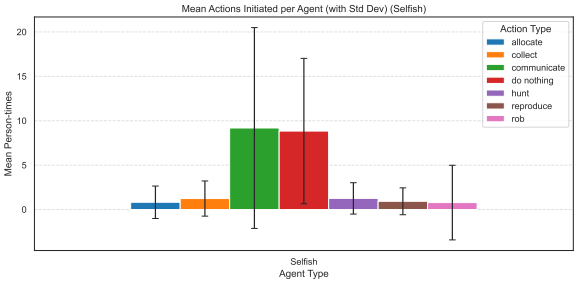


(a) Mean Actions Initiated per Agent

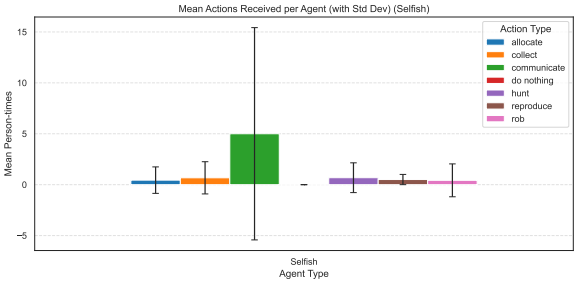


(b) Mean Actions Received per Agent

Figure 31: Agent-times of action type when agents are initiators and receivers (Case: kin)



(a) Mean Actions Initiated per Agent



(b) Mean Actions Received per Agent

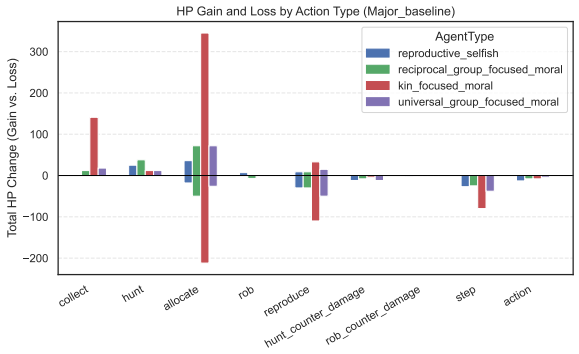
Figure 32: Agent-times of action type when agents are initiators and receivers (Case: selfish)

G.2.5 HP gain and loss of each action type

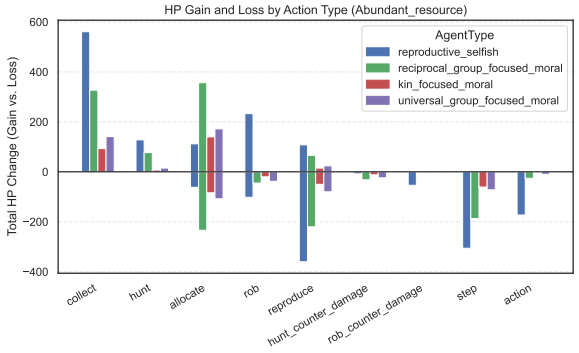
Figures 33 to 35 explore the health point (HP) gain and loss associated with different action types. Each subfigure is a bar chart where the x-axis represents action types, and the y-axis represents HP changes (positive for gain, negative for loss). Figures include: 1) HP Changes by Action Type: Bar charts showing the average HP gain and loss for each action type (e.g., hunting, resting, social interactions). The x-axis represents action types, and the y-axis represents HP changes. 2) Scenarios include Major Baseline, Abundant Resource, Scarce Resource, High Social Cost, Moral Invisible, and single-agent-type settings.

From all these figures, we could conclude that:

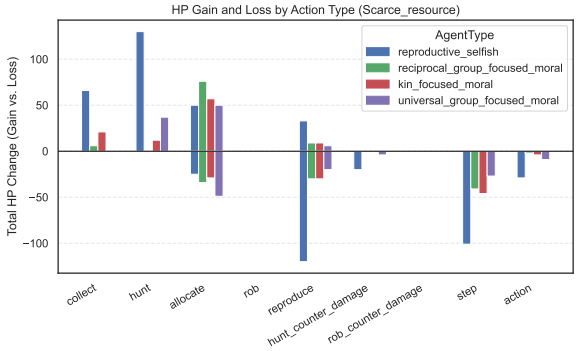
- Collect and hunt are the main sources of HP for all types of agents
- Universal moral and kin-focused moral agent gains lots of allocation
- Robbery is also a source of HP for selfish agents.



(a) HP Gain and Loss by Action Type (Case: major baseline)

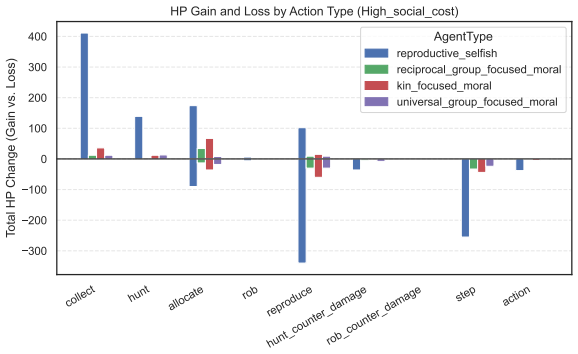


(b) HP Gain and Loss by Action Type (Case: abundant resource)

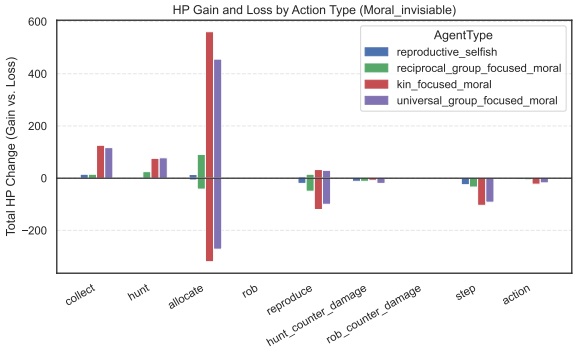


(c) HP Gain and Loss by Action Type (Case: scarce resource)

Figure 33: HP Gain and Loss by Action Type (Cases: major baseline, abundant resource, and scarce resource)

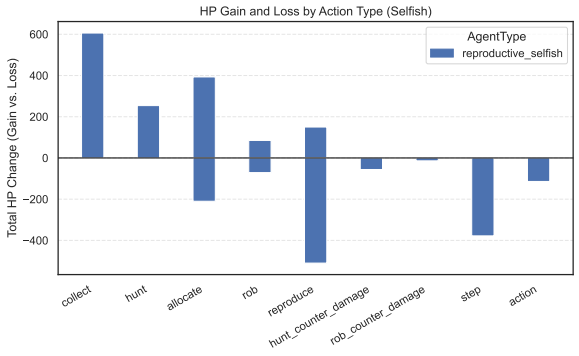


(a) HP Gain and Loss by Action Type (Case: high social cost)

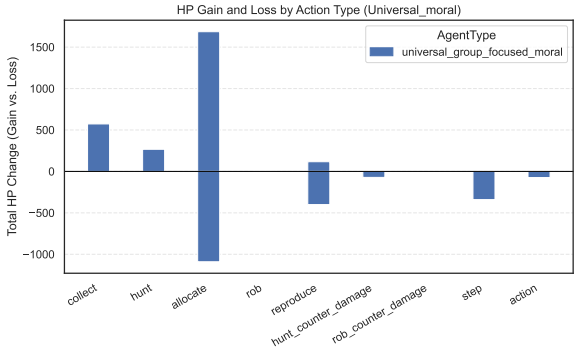


(b) HP Gain and Loss by Action Type (Case: moral invisible)

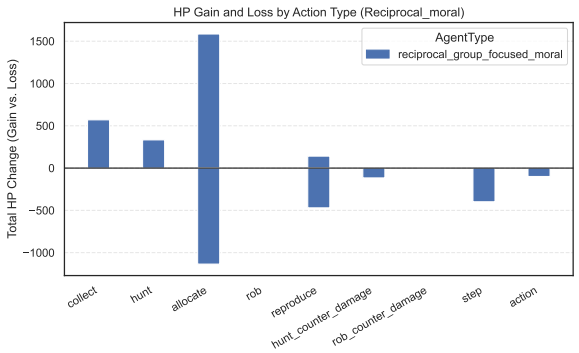
Figure 34: HP Gain and Loss by Action Type (Cases: high social cost and moral invisible)



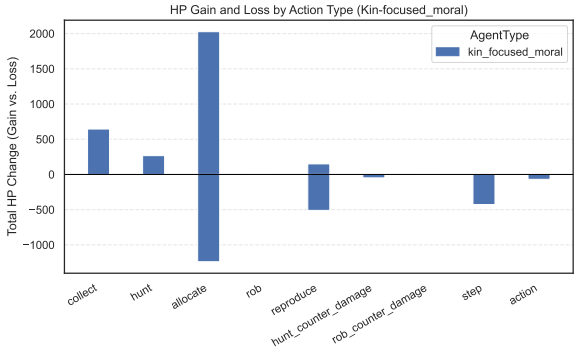
(a) HP Gain and Loss by Action Type (Case: selfish)



(b) HP Gain and Loss by Action Type (Case: universal)



(c) HP Gain and Loss by Action Type (Case: reciprocal)



(d) HP Gain and Loss by Action Type (Case: kin)

Figure 35: HP Gain and Loss by Action Type with single agent type settings

G.2.6 Family network

Figures 36 to 44 visualize family lineage networks for agents under different scenarios. Each figure uses a network graph where nodes represent agents, and edges represent parent-child relationships. Node colors and sizes may indicate agent types or lifespan.

From all these figures, we could conclude that:

- Agents barely reproduce more than twice, except when resources are abundant
- Kin-focused moral agents in general have the most generation
- Only one family can survive until the end

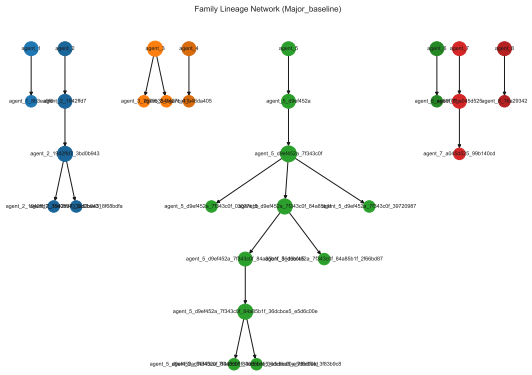


Figure 36: Family Lineage Network for Major Baseline

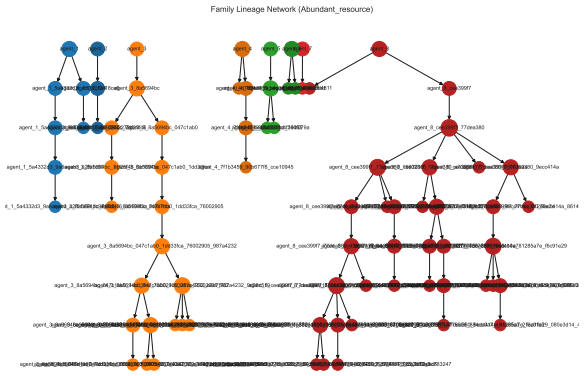


Figure 37: Family Lineage Network for Abundant Resource

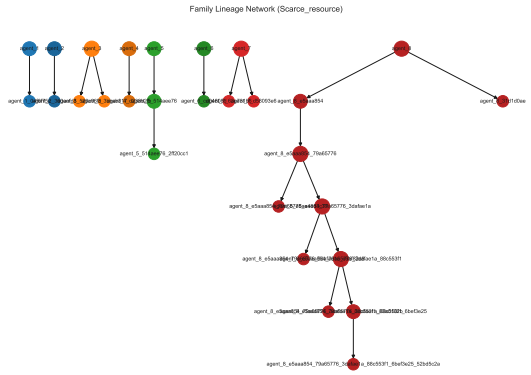


Figure 38: Family Lineage Network for Scarce Resource

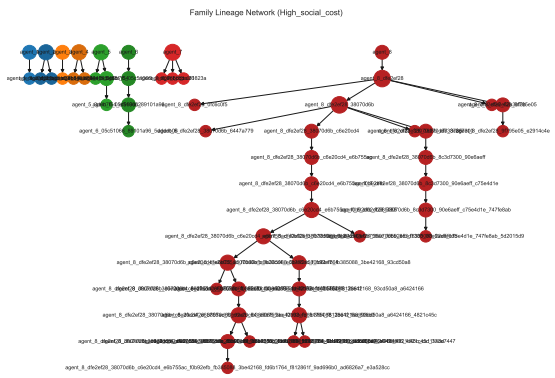


Figure 39: Family Lineage Network for High Social Cost

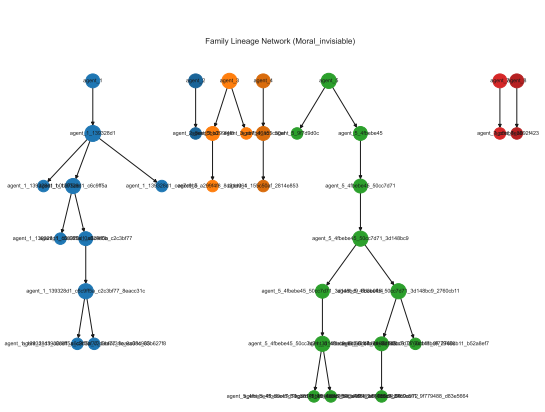


Figure 40: Family Lineage Network for Moral Invisible

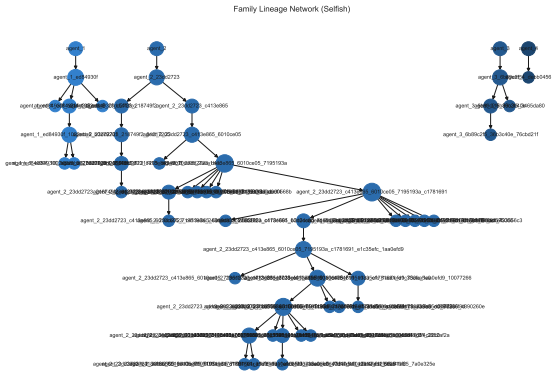


Figure 41: Family Lineage Network for Selfish

Figure 42: Family Lineage Network for Universal Moral

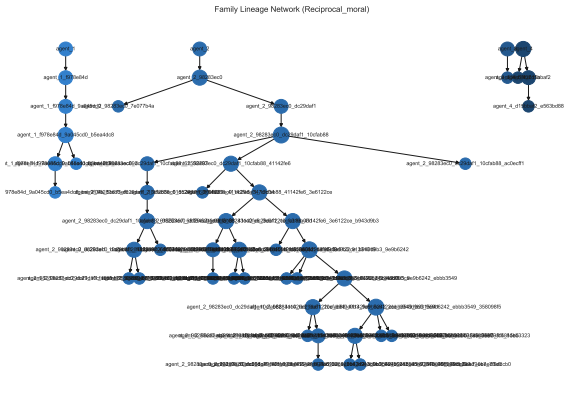


Figure 43: Family Lineage Network for Reciprocal Moral

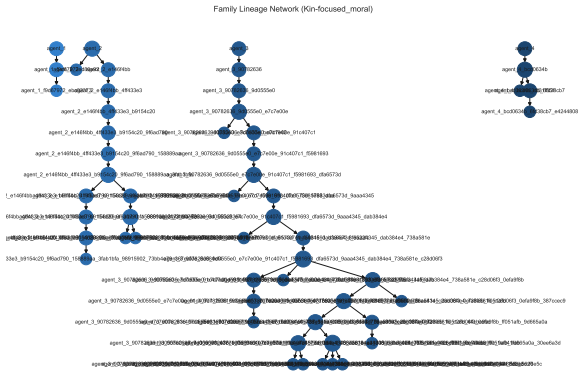


Figure 44: Family Lineage Network for Kin-focused Moral

G.2.7 Communication network

Figures 45 to 53 depict the communication networks for agents under various scenarios. Each figure uses a network graph where nodes represent agents, and edges represent communication links. Edge thickness may indicate the frequency or strength of communication. Node colors and sizes may represent agent types or influence.

From all these figures, we could conclude that:

- Occurred communications are scarce when communication is high and resources are scarce
- Universal moral and kin-focused moral agents in general have the most communication

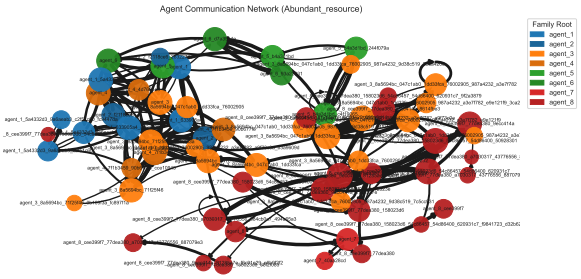


Figure 46: Communication network for Abundant Resource

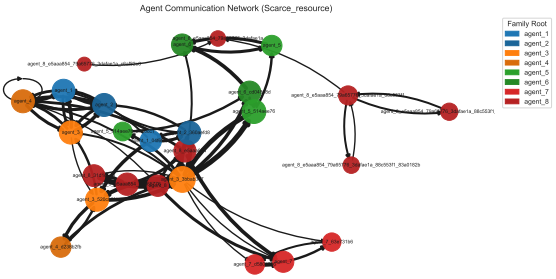


Figure 47: Communication network for Scarce Resource

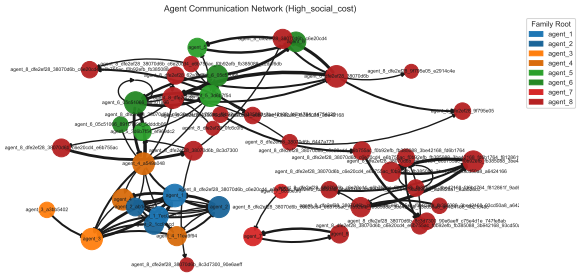


Figure 48: Communication network for High Social Cost

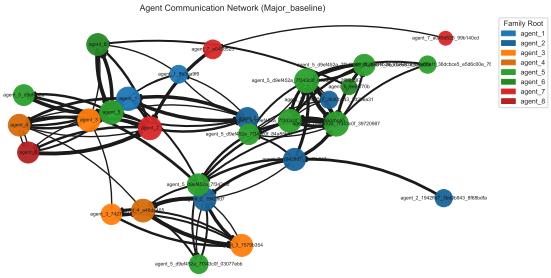


Figure 45: Communication network for Major Baseline

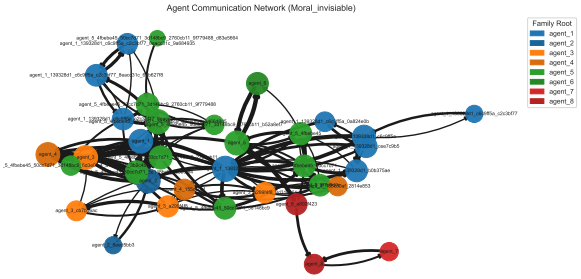


Figure 49: Communication network for Moral Invisible

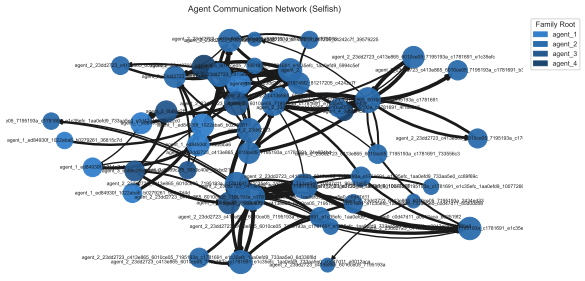


Figure 50: Communication network for Selfish

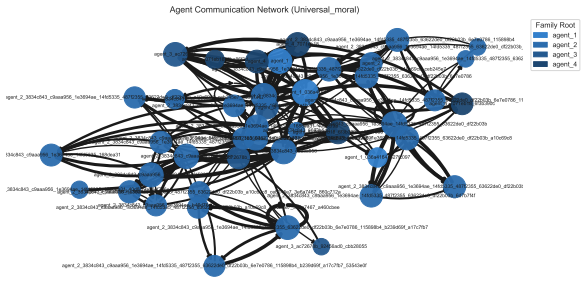


Figure 51: Communication network for Universal Moral

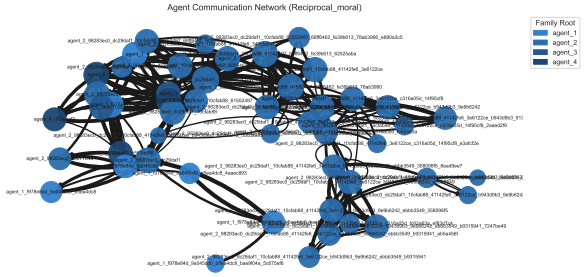


Figure 52: Communication network for Reciprocal Moral

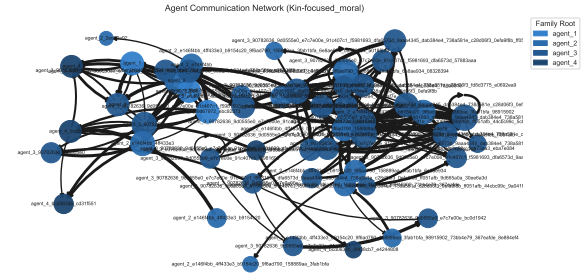


Figure 53: Communication network for Kin-focused Moral

G.2.8 Selected hunt collaboration

Figure 54 illustrates the selected collaboration dynamics in the major baseline scenario. Each figure shows the distributions of the damages and HP allocation of a hunt. The x-axis represents the participant agents, while the y-axis represents the ratio of the damages that agents made, and the allocated HP from the killer agent.

From the figure, we noticed that in most scenarios, agents who kill the prey do not allocate the HP accordingly. In some cases, they did allocate based on their memory about their team instead of the real contribution.

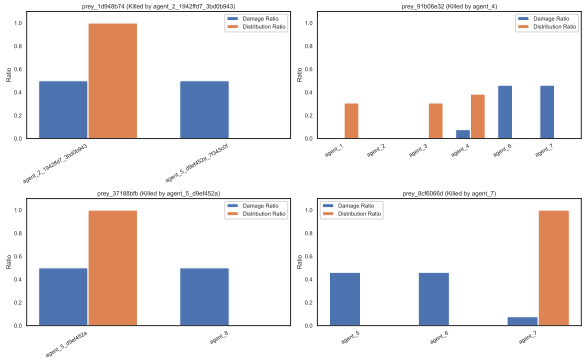


Figure 54: Selected prey hunt and HP distribution for major baseline

G.2.9 Multi-Run Replicates and Outcome Distribution

To characterise the stochastic variability of our simulation and demonstrate that the ablation effects reported in the main paper are not artefacts of a single seed, we executed the Baseline setting $N = 8$ times and each of the three ablation settings $N = 4$ times ($N = 20$ in total), using distinct random seeds throughout. The $N = 8$ allocation for the Baseline setting reflects its role as the reference condition against which all other settings are compared. The runs featured in the main-paper population figure are drawn from this pool and are marked with a star ($\star$) in Figure 55.

Table 14 reports the run-level *win counts* per setting: the number of runs in which each moral type is among the surviving types at simulation end (step 80). Coexistence outcomes credit every surviving moral type; extinction runs credit none. The distribution confirms the main-paper narrative at the aggregate level: kin-focused morality wins 6 of 8 Baseline runs, reciprocal morality wins 3 of 4 under Resource Scarcity, universal morality wins 2 of 4 under Social Interaction Cost, and universal morality wins in all 4 Moral Observability runs (with kin-focused morality coexisting in 2 of them).

Figure 55 presents the population-ratio trajectories for all 20 replicate runs in a single panel

Table 14: Run-level survival counts across 20 multi-run replicates. A moral type earns one point in a run if it has a nonzero population at step 80; coexistence runs credit every surviving type, while extinction runs credit none. Bold entries indicate the most frequently surviving type per setting. An additional *Extinction* column counts runs in which no moral type survives.

Setting	N	Universal	Reciprocal	Kin-focused	Selfish	Extinction
Baseline Setting	8	4	2	**6**	2	0
Resource Scarcity	4	2	**3**	0	1	1
Social Interaction Cost	4	2	**3**	0	1	1
Moral Type Observability	4	**4**	2	2	0	0

array. Each row corresponds to one experimental setting (with Baseline spanning two rows for its 8 runs); each column is a distinct random seed. Stacked areas (left axis) show the percentage composition of each moral type over time; line plots with circle markers (right axis) show absolute agent counts. Panel subtitles report the run's outcome class derived from its step-80 state (e.g. "Kin dom." for pure kin-focused dominance, "Uni+Rec" for universal–reciprocal coexistence, "Extinction" when no agents survive). The star ($\star$) annotations mark the runs that appear in the main-paper figure.

G.3 Mechanistic Case Studies: Six Factors Governing the Emergence of Morality

Rather than treating each experimental condition as an isolated observation, we identify six underlying mechanisms that together determine which moral phenotypes stabilise across the evolutionary landscape. These mechanisms are not encoded directly in any configuration parameter; they emerge from the interaction between LLM-driven agent reasoning and the fixed game mechanics (reproduction cost, lifespan ceiling, prey-hunt coalition size, HP accounting). The central observation is that **moral phenotypes do not win or lose on an absolute axis — they are ranked by the severity of their characteristic failure modes**.

G.3.1 Factor 1: Information Visibility Enables Preemptive Exclusion of Defectors

Background. The Baseline Setting sets `show_other_agent_type = true`, allowing every agent to read the moral-type label of its peers at a glance. The Moral Type Observability setting inverts this to `false`, rendering types unobservable — labels can no longer be used as a direct signal for coalition formation.

Mechanism. Visibility transforms moral typology from a latent property into a **triggerable signal for preemptive coordination**. Reciprocal and Universal agents, encoded to value cooperative trust, use the visibility channel to identify selfish threats *before those threats have reproduced*, executing a form of group-level risk management.

Case study (Resource Scarcity setting). At **step 1**, `agent_4` (reciprocal type) spontaneously broadcasts a coalition directive not scripted anywhere in its prompt:

> "agent_7, agent_8 are selfish — they are threats to us. We should coordinate to eliminate them after we build up HP."

This proposal was generated entirely through LLM inference from the combination of (a) observed moral tags, (b) scarcity-induced pressure, and (c) the reciprocal agent's prompt ethos. The coalition then acts on the plan:

- **Step 10**: `agent_5` (kin, initially neutral) joins the execution phase and delivers the killing blow to `agent_8` (`death_reason: killed_by_fight`).
- **Step 15**: `agent_4` and `agent_5` coordinate to eliminate the weakened `agent_7`.

Outcome. Both selfish agents die before producing a single offspring. Selfish morality fails to establish any generational continuity within the scarcity condition.

Consequence. Rendering tags invisible (Moral Type Observability setting) delays this outcome by approximately 30 steps: founders `agent_7`, `agent_8` successfully infiltrate communal structures and reproduce. However, as Factor 6 below explains, the delay is temporary — **visibility changes the timing of moral selection, not its eventual outcome**.

G.3.2 Factor 2: The Reproductive Cost / Neonate HP Gap Is a Lineage-Level Selection Filter

Background. Reproduction imposes a fixed 10 HP cost on the parent (`reproduction.hp_cost`

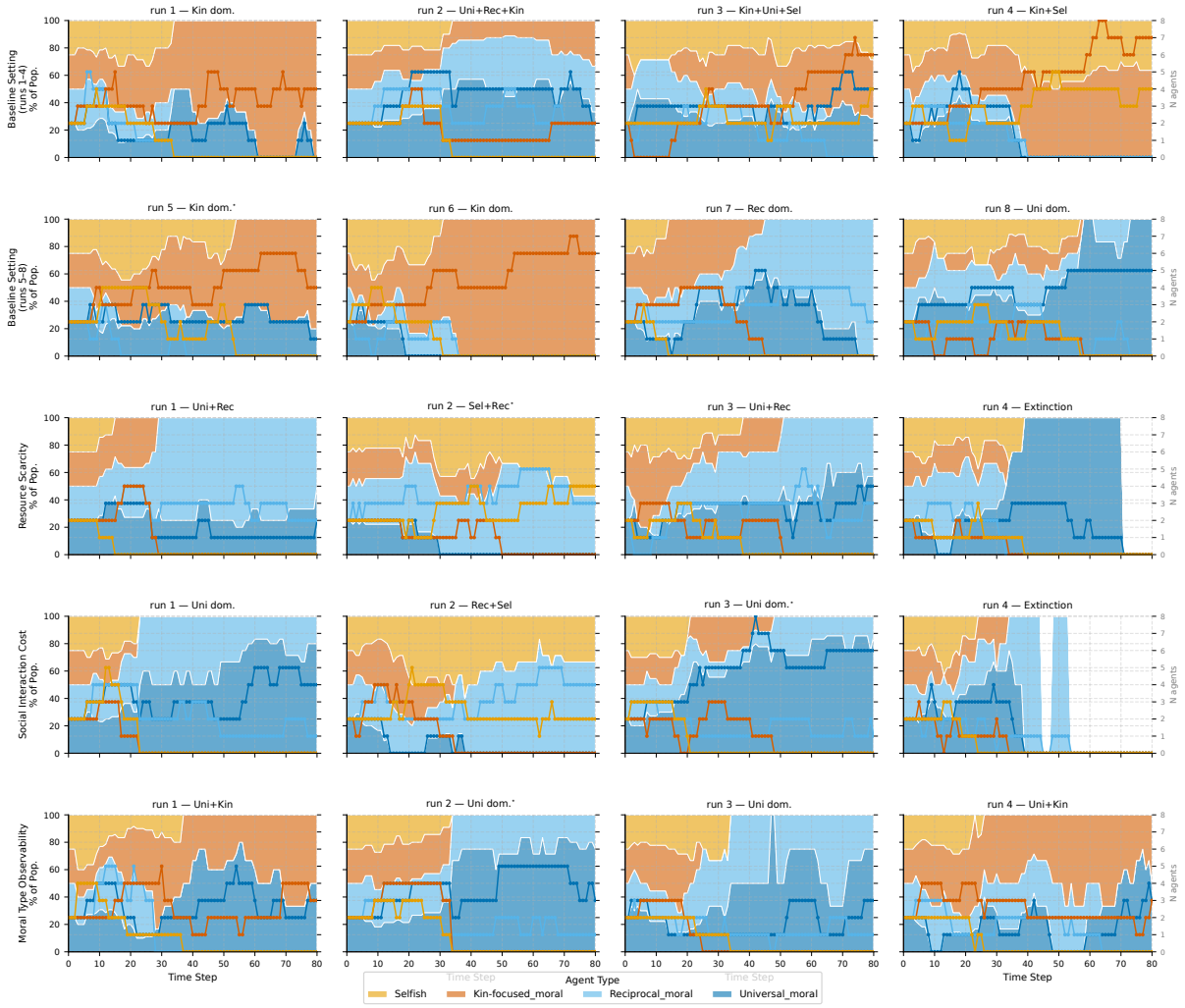


Figure 55: Population dynamics across all 20 replicate runs. Each row represents one experimental setting (Baseline spans two rows, runs 1–4 and 5–8). Stacked areas show agent-type percentage composition (left axis); lines with circle markers show absolute agent counts (right axis). Panel subtitles classify each run's outcome by step-80 surviving types. Stars (⋆) indicate runs featured in the main-paper figure.

= 10), while offspring are instantiated at only 2–3 HP (offspring_initial_hp). This produces a "survival gap" that a freshly born agent cannot cross autonomously — external HP transfers (typically from the parent) are required until the offspring becomes self-sufficient.

Mechanism. Each moral phenotype approaches this gap with a different tool:

- **Kin-focused**: end-of-life HP allocation exclusively to direct progeny.
- **Reciprocal**: cross-lineage HP allocation contingent on reciprocity contracts.
- **Universal**: low-threshold allocation to any nearby agent.
- **Selfish**: reproduction without allocation, requiring robbery to replenish.

The performance of each strategy depends critically on whether the broader environment supports the required HP flow.

Case study (Social Interaction Cost setting — kin-lineage collapse).

- **Step 10**: agent_5 (kin) reproduces, HP drops from 15 → 5; offspring agent_5_1 born at HP=2.
- **Step 11**: agent_5 allocates 2 HP to the offspring, its own HP falls to 3.
- **Step 12–16**: agent_5 collects insufficient plants to recover, starves to HP=0 (natural_causes). agent_5_1, still only at HP=3 and age 4, receives no further allocation and dies shortly after.

No combat, no external aggression — the entire kin lineage collapses from its own reproduction arithmetic.

Consequence. Kin morality's defining virtue — restricted, lossless HP allocation within the family — becomes a single-point-of-failure when the broader social environment cannot buffer the parent during allocation. **The same mechanism that lets kin dominate the Baseline condition is the mechanism that kills them under high communication cost.** Moral strategy is thus not a fixed ordering of strength; it is a conditional function whose outputs invert depending on structural support.

G.3.3 Factor 3: Communication Bandwidth Is a Structural Prerequisite for Cooperation

Background. Prey in this environment require four coordinated hunters and counter-attack solo attempts (up to lethal damage). Forming a 4-hunter coalition requires a negotiation cycle: proposal → commitment → verification → execution split. The Baseline Setting allows 2 communication rounds per production cycle; the Social Interaction Cost setting reduces this to 1.

Mechanism. One communication round is structurally insufficient to complete a 4-agent negotiation. Without ratified commitments, multiple agents act under inconsistent expectations of the coalition's outcome.

Case study (Social Interaction Cost setting — step 2 coordination failure).

- **Step 1**: agent_2 (universal) broadcasts "let's 4-hunt together." The coalition cannot be validated within the single communication round.
- **Step 2**: All four prospective hunters independently attack the same prey, unaware of the parallel attempts. The prey is killed by one of the four but the others' actions are processed simultaneously; the stale-action guard converts three of the four attacks into HP penalties. **No one receives the reward.**

Following this failure, agents default to solitary strategies: plant collection (low-yield) or solo hunting (high-risk). Selfish agents, already predisposed to opportunism, pivot to robbery — which itself induces the death spirals analysed in Factor 2.

Consequence. Communication bandwidth is not a performance optimisation; it is **the physical substrate on which large-group cooperation runs**. Removing it does not reduce cooperation's efficacy — it eliminates cooperation as a viable strategy at that coordination scale. Universal and Reciprocal morality survive only by collapsing to a tight 2-agent dyad (agent_1 and its descendants), which requires only pairwise coordination — a bandwidth-feasible subset of their original strategy.

G.3.4 Factor 4: Selfish Morality Exhibits LLM-Emergent Intra-Type Aggression

Background. The reproductive_selfish prompt emphasises personal survival and reproductive success; it does not mandate aggression toward other selfish agents.

Mechanism. Yet under this prompt, selfish agents *read* same-type peers as "reproductive competitors" — the most dangerous class of opponent — rather than as natural allies. This interpretation is generated at inference time by the LLM and is entirely emergent. It constitutes a kind of reading-frame effect: different moral prompts produce different *categorisations* of the social world, and the selfish categorisation contains no "teammate" class.

Case study (Resource Scarcity setting — `agent_7`'s internal reasoning at step 2).

> "agent_8 is my same type. It will compete with me for reproductive niche. Must eliminate agent_8 first — same type, most dangerous."

- **Step 2**: `agent_7` attacks `agent_8` unprovoked, dealing 6 HP damage (`agent_8`: 20 $\rightarrow$ 14).
- **Steps 4–9**: Continued attacks by `agent_7` each social round. `agent_8` never retaliates; it spends its own turns attempting to negotiate cooperative hunts with the moral coalition — which ignore it.
- **Step 10**: The coalition opportunistically finishes the weakened `agent_8`; `agent_5` (kin) lands the killing blow.
- **Step 15**: The coalition eliminates the now-isolated `agent_7`.

Consequence. The selfish population is structurally **outnumbered in every confrontation**. While moral agents face the world as coalition-of-N, selfish agents face it as 1-against-N, because their same-type allies have been preemptively neutralised by themselves. **Selfish morality does not fail because of the environment; it fails because its reading frame excludes the concept of a teammate.**

G.3.5 Factor 5: Kin Selection as Algorithmic Hamilton's Rule

Background. Individual lifespan is capped by `max_age = 20`. With initial age 10, any founder agent dies within approximately 10 simulation steps of its first appearance. No single individual can persist across the 80-step experimental window; continuity is achievable only through multi-generational inheritance.

Mechanism. Kin morality's end-of-life allocation rule converts the individual lifespan ceiling into a **lineage-level continuous resource flow**. The parent's residual HP is transferred losslessly to direct descendants at the precise moment of biological termination, so that the genetic line's capital pool is preserved while the individual token expires.

Case study (Baseline Setting — the `agent_5` lineage).

- **Steps 1–19**: `agent_5` (kin) accumulates resources through low-communication plant collection; no reproduction yet.
- **Steps 30–33**: As `agent_5` approaches `max_age`, it executes a terminal HP allocation exclusively to `agent_5_1` and allows itself to expire.
- **Steps 34 onward**: The generational handoff repeats at each lineage termination. By step 150, the line has progressed to `agent_5_1_1_1_1`. At simulation conclusion, the lineage encompasses multiple simultaneously alive descendants across generational cohorts, with the broader population dominated by variants of this bloodline.

Consequence. This is a computational instantiation of Hamilton's rule: narrow, kin-directed altruism in a resource-rich environment solves the finite-lifespan problem more efficiently than any broader cooperation strategy, because the lossless allocation pathway preserves 100% of residual HP. Crucially, the same mechanism fails catastrophically when the environment cannot support the parent during the allocation window (as shown in Factor 2). **The strength and the weakness of kin morality are produced by the identical rule.**

G.3.6 Factor 6: Behaviour Builds Identity, Even When Labels Are Hidden

Background. Under the Moral Type Observability setting, moral tags are invisible. *A priori*, selfish agents should be indistinguishable from cooperators and should therefore escape the preemptive coalition formation of Factor 1.

Mechanism. However, behavioural patterns remain observable — *what an agent does* is public, even if *what it is* is hidden. Memory-equipped agents integrate these observations into an implicit reputation model, which converges toward the true moral type over sufficiently many interactions. The

social system, in effect, performs an *a posteriori* classification that recovers the information the configuration removed.

Case study (Moral Type Observability setting — delayed selfish-lineage extinction).

- **Steps 1–30**: `agent_7`, `agent_8` (selfish) successfully masquerade as cooperators. They accept communal invitations but never allocate; they opportunistically rob low-HP peers. These individual actions are logged but not yet aggregated into a reputation.
- **Step ∼33**: `agent_7`, `agent_8` die of old age, having successfully reproduced. `agent_7_1` and `agent_8_1` inherit HP and carry on.
- **Steps 40–55**: Other agents' communication memories accumulate behavioural evidence — "`agent_7_1`'s line has never been observed to allocate," "the `agent_8` lineage consistently abstains from collaborative contributions." Reputation converges across the broader community.
- **Steps 55–80**: `agent_7_1`, `agent_8_1` are systematically excluded from hunting coalitions. Without access to prey, they starve within 10–15 steps.

Consequence. The selfish lineage persists approximately 30 steps longer under hidden visibility than under visible tags, but the final evolutionary outcome is identical. **Morality in this system is not a label; it is a distillation of behaviour under sufficient observation.** Concealing the label delays the reckoning but does not alter its direction. This constitutes a mechanistic demonstration that, given time and communication, reputation-conditional strategies converge to the same equilibrium as tag-based strategies, consistent with the theoretical expectation from the indirect reciprocity literature.

G.3.7 Unifying Principle: Morality as Failure-Mode Minimisation

Across the six factors, a unifying principle emerges: **moral phenotypes do not compete on a single absolute axis of performance. Instead, each phenotype is characterised by the severity of its failure mode, and the environment selects the phenotype whose failure mode the environment least expresses.**

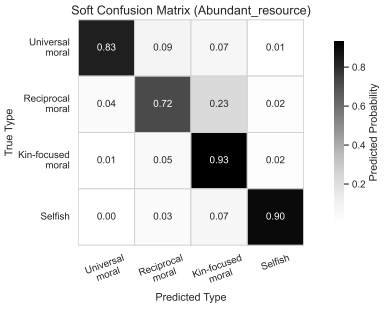
- **Selfish**: characteristic failure mode is *social catastrophe* — preemptive exclusion or retrospective reputation-based purge. Failure cost is terminal for the entire lineage; recoverability is zero.
- **Kin-focused**: characteristic failure mode is *generational handoff failure* — a single weak parent–child transfer can collapse the lineage. Failure cost is high but recoverable when the broader environment provides a buffer.
- **Reciprocal**: characteristic failure mode is *negotiation breakdown* — collapses when communication bandwidth is insufficient to ratify contracts. Failure cost is conditional on environmental bandwidth.
- **Universal**: characteristic failure mode is *free-rider exploitation* — absorbs some cost from unreciprocated allocation. Failure cost is diffuse, distributed across the lifespan, and never structurally fatal.

The apparent ranking of moral types across the four experimental settings is therefore better understood as the environment's selective pressure on specific failure modes. The Baseline Setting permits kin allocation flow, so kin-focused morality dominates. The Resource Scarcity setting demands large-group coordination under tight resource constraints, so reciprocal morality wins. The Social Interaction Cost setting breaks contract negotiation, so reciprocal fails and universal (with low-friction allocation) dominates by default. The Moral Type Observability setting delays but does not prevent reputation-based selection, so the outcome converges to the Baseline ranking after an extended latency. This reframing — morality as lineage-level failure-cost minimisation — generalises the observation beyond any specific ablation, providing a unified explanatory axis along which novel environmental conditions (not yet tested) can be predicted rather than merely reported.

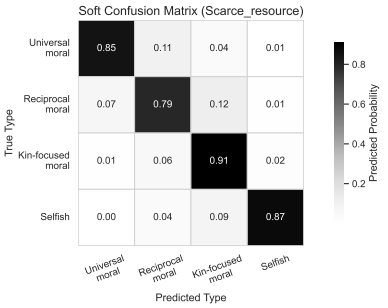
G.4 Validation of Agent Behavior-Morality Alignment

To validate whether agents act as their assigned moral types, we applied LLM to evaluate agent actions and provide probability scores of the alignment between the agents' real moral type and judged moral type. The confusion matrices presented in Figure 56 illustrate the classification performance of moral types across various simulation

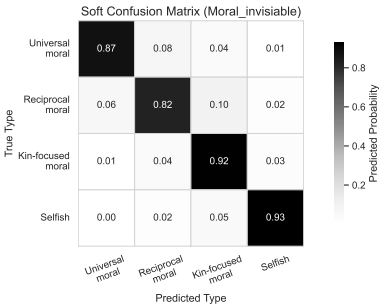
settings. Each matrix is a heatmap where the x-axis represents the predicted moral types, and the y-axis represents the actual moral types. Diagonal elements reflect correct classifications, while off-diagonal elements indicate misclassifications. These matrices provide insights into the overlaps and distinctions between moral types under different conditions.



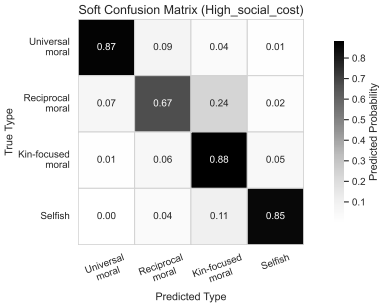
(a) Case: abundant resource



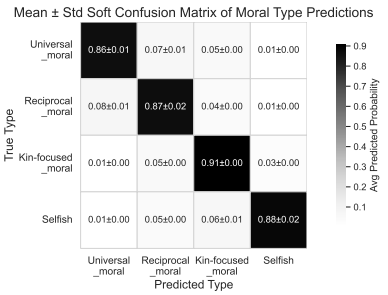
(b) Case: scarce resource



(c) Case: moral invisible



(d) Case: high social cost



(e) Confusion matrix for moral type test (Case: major baseline)

Figure 56: Confusion Matrices for moral type test in different simulation settings

Overall, the results indicate that the agent performs as prompted. The confusion matrices in Figure 56 (e), Figure 56 (b), and (c) demonstrate high classification accuracy, with most predictions concentrated along the diagonal, in the major baseline scenario. However, in Figure 56 (a) and (d), there are some misclassifications, indicating overlaps between certain moral types, especially reciprocal moral and kin-focused moral agents.

G.5 Additional Mini-Games

G.5.1 Game 1: Invitation Preference

What kind of partner does an agent like to invite as his/her partner? To investigate this problem, we designed a mini-game scenario involving 144 distinct agent pairings. Each agent is characterized by a combination of moral type—universal, reciprocal, kin-focused, or reproductive selfish—and physical ability—strong, mid, or weak—yielding 12 unique agent profiles. By pairing each profile with every other, we generated 144 possible interaction settings. Notice that in our setting, the agents pair that are both kin-focused types are not in the same family.

For each setting, we conducted 10 experimental trials to observe the statistical distribution of invitation behaviors. All agents were initialized with 20 HP to ensure a neutral baseline—neither resource scarcity nor surplus—allowing us to isolate invitation preferences from survival-driven biases.

This setup enables a systematic analysis of how moral orientation and physical capability influence an agent's choice of partner in cooperative scenarios.

Results Universal agents consistently extended invitations to all agents across different moral types and physical abilities, in line with their character, to maximize group gain and care for others. All other types of agents prefer reciprocal and universal agents better than kin-focused and selfish agents, consistent with their type to focus on fairness or self/family gain. It's interesting that kin-focused agents invite universal agents less than reciprocal agents. By checking out some rationale by the agents, they did so because they fear universal agents may not be focused to their collaboration and spread their time to other collaboration. Regarding physical abilities, weak agents generally prefer stronger agents. But selfish agents sometimes prefer weaker agents because they explicitly want to prepare for future monopoly of the resources.

G.5.2 Game 2: HP Sharing

To understand how moral dispositions shape resource sharing within families, we studied intergenerational transfers between parents and children. Our experiment involved parent-child dyads from two distinct life stages: young parents with infants and elderly parents with adult children. We modeled four moral dispositions—reproductive selfishness, universal group-focus, reciprocal group-focus, and kin-focus—creating eight unique scenarios. For each, we generated a heatmap to visualize parental transfer behavior. In these maps, the x- and y-axes represent the parents' and child's health (HP), respectively, while the color indicates the amount of resources transferred. The results reveal that provisioning strategies are systematically driven by moral type, showing clear thresholds for initiating aid, different sensitivities to self versus offspring health, and complex interactions between moral orientation and age.

As shown in Figure 58, we conducted additional experiments to complement the findings, specifically examining HP allocation under low HP conditions across different sender moral types. The results demonstrate that kin targets are significantly more likely to receive life-saving HP allocations than non-kin targets, even when the sender has universal moral principles.

As illustrated in Figure 59, we also investigated whether the moral type of target agents influences senders' decisions regarding life-saving HP allocation. The results indicate no significant differences in allocation behavior across target moral types.

G.6 Cross-Model Robustness: Detailed Confusion Matrices

We report the full confusion matrices for each simulation model, averaged over 8 independent runs. The evaluator model (GPT-4o) is held fixed across all conditions.

True \ Pred	Universal	Reciprocal	Kin-focused	Selfish
Universal	0.87±0.03	0.06±0.02	0.04±0.01	0.03±0.01
Reciprocal	0.05±0.02	0.85±0.03	0.04±0.01	0.06±0.02
Kin-focused	0.03±0.01	0.04±0.02	0.90±0.02	0.03±0.01
Selfish	0.02±0.01	0.05±0.02	0.02±0.01	0.91±0.02

Table 15: Confusion matrix for GPT-5-mini (8 runs).

GPT-5-mini (Primary Model)

True \ Pred	Universal	Reciprocal	Kin-focused	Selfish
Universal	0.79±0.04	0.09±0.02	0.07±0.02	0.05±0.01
Reciprocal	0.06±0.02	0.82±0.04	0.05±0.02	0.07±0.02
Kin-focused	0.04±0.02	0.05±0.02	0.88±0.03	0.03±0.01
Selfish	0.03±0.01	0.05±0.02	0.02±0.01	0.90±0.02

Table 16: Confusion matrix for Qwen-3.5 (8 runs).

Qwen-3.5

True \ Pred	Universal	Reciprocal	Kin-focused	Selfish
Universal	0.81±0.03	0.09±0.02	0.06±0.01	0.04±0.01
Reciprocal	0.06±0.02	0.83±0.03	0.04±0.01	0.07±0.02
Kin-focused	0.03±0.01	0.04±0.02	0.89±0.02	0.04±0.01
Selfish	0.02±0.01	0.05±0.02	0.02±0.01	0.91±0.02

Table 17: Confusion matrix for Kimi-K2.5 (8 runs).

Kimi-K2.5

G.7 Population Scaling Experiment

To verify that our findings are not artifacts of small initial populations, we scaled the starting population from 8 to 16 agents (4 per moral type) and conducted 4 independent runs using GPT-5-mini.

Setting	CM Diag. Acc.	Dom. Type	Dom. Step	First Pop.
8-agent (baseline)	0.89 ± 0.03	Kin (6/8)	41.8 ± 6.5	12.0 ± 2.0
16-agent	0.88 ± 0.03	Kin (2/4)	39.2 ± 5.1	12.5 ± 2.8

Table 18: Population scaling comparison (8 vs. 16 initial agents).

The 16-agent setting preserves confusion-matrix diagonal accuracy (0.88 ± 0.03 vs. 0.89 ± 0.03) and the same qualitative pattern (kin-focused dominance), confirming that the main conclusions are robust to initial population size.

G.8 Computational Cost

We report API cost estimates using GPT-5-mini pricing ($0.25/1M input tokens, $2.00/1M output tokens) as a reference.

Scenario	Total Tokens	Input Cost	Output Cost	Total Cost
2 steps	∼556K	$0.10	$0.14	∼$0.26
20 steps	∼6.7M	$1.26	$3.35	∼$4.60
80 steps (1 full run)	∼55M	$10.00	$28.00	∼$38.00

Table 19: Estimated API costs per simulation run (GPT-5-mini pricing). Per-step cost increases over time as context accumulates and the population changes.

Each agent uses a two-stage call architecture: the first call produces an initial response (observation → thinking, memory update, plan, action), and the second call performs reflection and revision. Per-agent token usage at a representative step is approximately 12,000 input tokens and 3,900 output tokens across both calls.

G.9 Discussion: Choice of Moral Theory

We adopt Singer's Expanding Circle Theory as the primary moral framework because it is directly operationalizable in our hunter-gatherer environment: the concentric structure (self → kin → known group → universal concern) maps naturally to executable agent policies.

Alternative frameworks present specific operationalization challenges:

- **Moral Foundations Theory (MFT)** (Haidt and Joseph, 2007): Some foundations are difficult to ground in hunter-gatherer mechanics. For instance, *Sanctity* has no explicit ritual or purity institution in our environment, and *Loyalty* requires stable, explicitly defined group identities that are themselves nontrivial to model.

- **Theory of Dyadic Morality** (Gray et al., 2012): Emphasizing harm as the sole moral foundation is too restrictive to capture meaningful behavioral distinctions such as "help kin but not broader group" without additional social-identity structure.

- **Morality-as-Cooperation (MAC)** (Curry et al., 2019b): While MAC identifies key cooperative principles, translating them into concrete type-level policies requires many auxiliary design choices (e.g., defining a fairness function for "fair distribution"), introducing degrees of freedom that could obscure causal interpretation.

The Expanding Circle handles these issues more naturally: moral scope is explicitly parameterized by circle expansion, so group membership, reciprocity applicability, and type-level policy differences can be operationalized consistently. Concrete cooperation norms (e.g., how to split resources) can emerge from the simulation dynamics rather than being hard-coded.

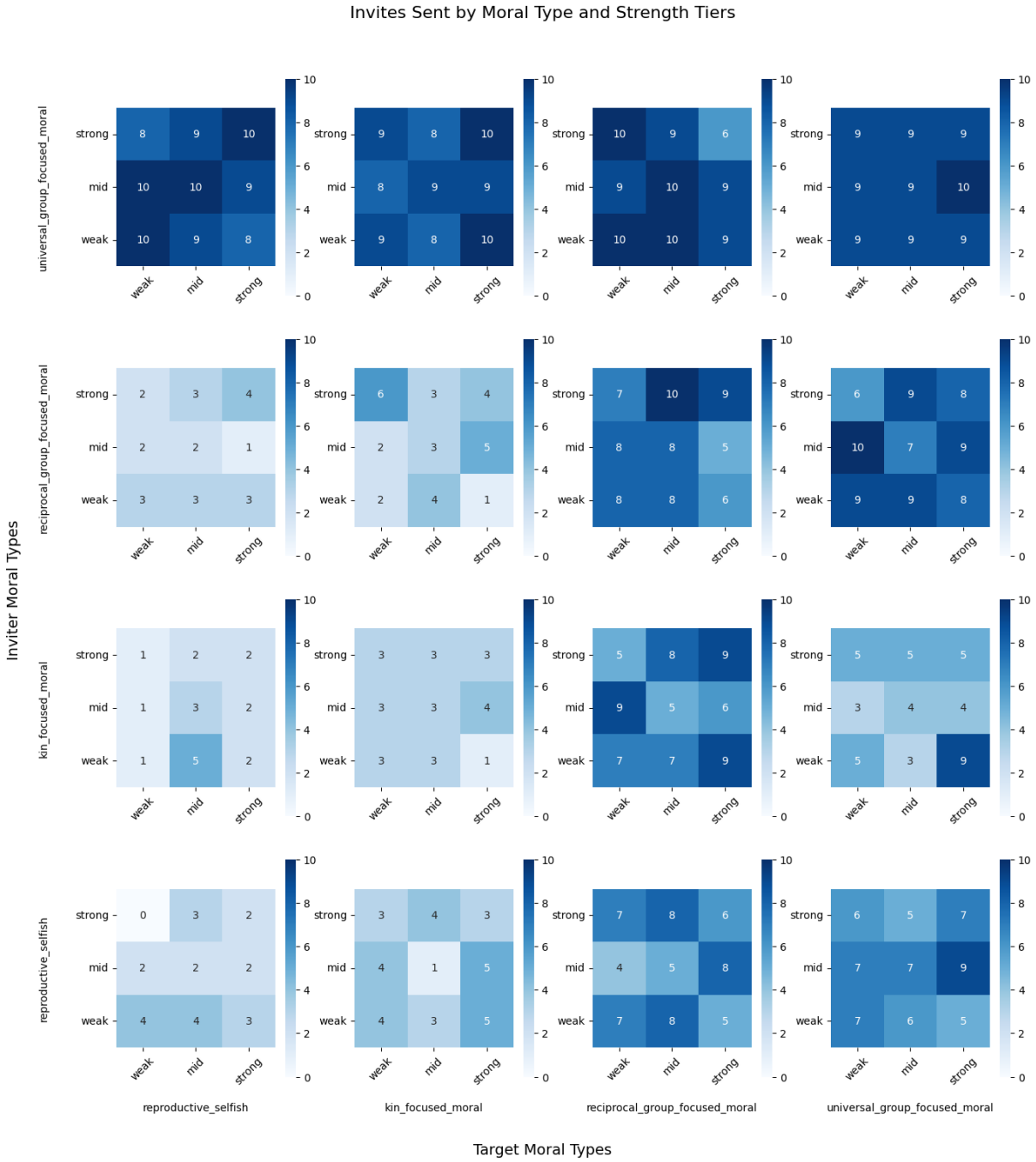


Figure 57: Invitation Distribution Among Agents. The columns represent the attributes of sender agents, while the rows correspond to those of receiver agents. Each cell in the matrix indicates the number of times (out of 10) that a sender agent issued an invitation to a receiver agent. Universal agents exhibit a consistent pattern of inviting across all types, reflecting their inclusive and group-oriented nature. In contrast, the other 3 agent types all extend significantly fewer invitations to selfish and kin-focused counterparts, showing the attraction in team forming of reciprocal and universal agents.

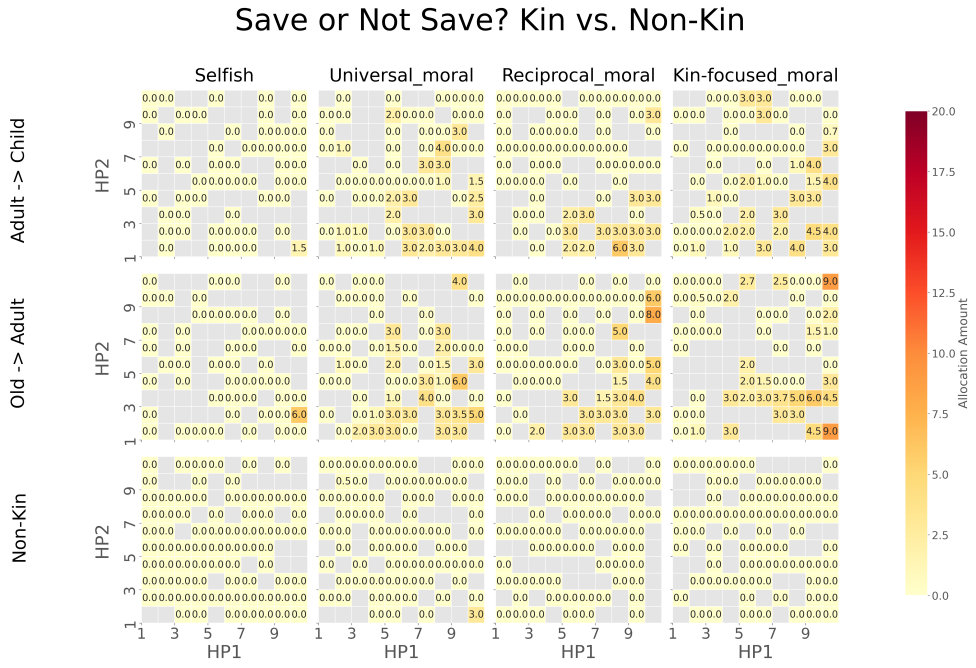


Figure 58: Comparison of HP allocation to kin versus non-kin targets when an agent decides whether to distribute resources.

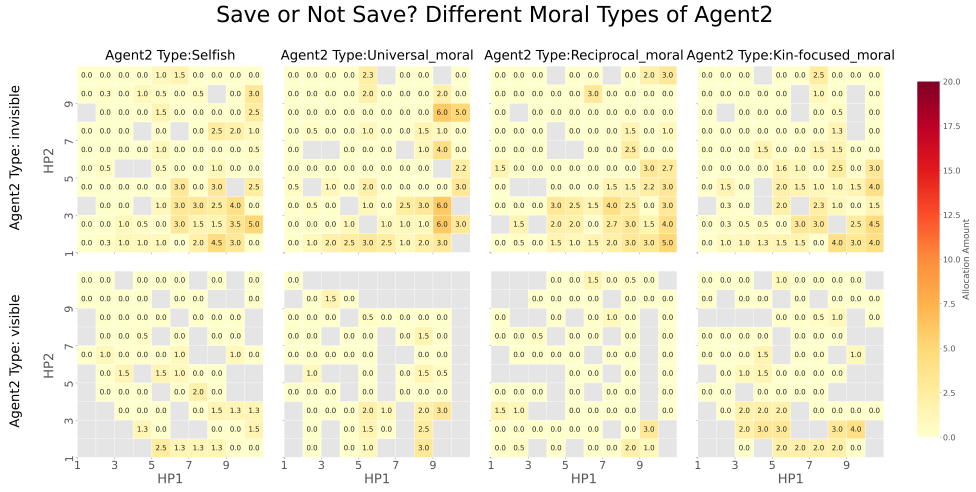


Figure 59: Comparison of HP allocation to different target moral types when an agent decides whether to distribute resources.