Dynamic Greed in Artificial Moral Agents

Nathaniel Kremer-Herman
Seattle University
Seattle, WA (USA)
nkh@seattleu.edu

Chengfei Jiang
Seattle University
Seattle, WA (USA)
cjiang@seattleu.edu

***Abstract*—With artificially intelligent agents on the rise, it is clear that autonomous decisionmakers should behave ethically and cooperatively. Multi-agent systems contain agents with differing agendas. A purely selfish agent will exploit others. A purely selfless one will be exploited. In multi-agent systems, each agent should adapt their degree of greediness versus cooperativeness according to their circumstances. We demonstrate that explicit moral reasoning allows agents to find this balance. We extend the Sugarscape agent-based model to showcase dynamic agent greed in a multi-agent system. Sugarscape societies with dynamic greed provide up to 31.5% higher societal survival, up to 2× the population, with agents who can live 1.7× as long, and who are up to 1.6× happier compared to their static counterparts. Agent greed moves toward a balance between selfishness and selflessness. We show that this balance promotes cooperation which is necessary for success.**

***Index Terms*—agent-based modeling, machine ethics, computational cultural modeling, cooperative systems.**

I. Introduction

As artificially intelligent (AI) agents continue to emerge and be widely deployed to make decisions on behalf of humans, ensuring that their behavior remains ethical is becoming increasingly important. It is no longer sufficient to merely rely on autonomous systems to perform tasks efficiently. They must also demonstrate ethical consideration, adaptability, and the ability to collaborate in order to make responsible decisions in dynamic environments.

From a computational perspective, greediness plays a crucial role in optimizing limited resources and adapting to environments with incomplete information. Greedy behavior aims at maximizing immediate gains. In multi-agent systems, agents are often designed to make decisions based on their self-interest (i.e. accomplishing their mission). For example, a self-driving car might choose the fastest route and fastest speed to minimize its travel time without considering how this choice affects overall traffic patterns or the safety of others.

It would be preferable for agents to balance accomplishing their missions with the potential for harm caused by their actions. Operationalizing an agent's level of greed provides a key mechanism for shifting self-interested behavior toward cooperative outcomes. When considerate behavior is needed, an AI agent should be able to adapt their degree of greediness (perhaps only momentarily) to the situation. We offer a computational perspective on greed for the purpose of designing AI agents that balance self-interest with moral principles.

We define AI agents as autonomous, goal-oriented, potentially self-improving computer processes, software objects, or algorithms which interact with an environment. Terminology in the artificial intelligence domain regarding agents is changing rapidly, but we liberally accept AI agents displaying a variety of intelligent behaviors. We focus on embedding explicit moral reasoning within agents. Our topics of discussion apply throughout this broadly defined class of AI agents.

Machine ethics is an emerging field that aims to embed AI systems with ethical reasoning capabilities, allowing them to navigate dilemmas and make morally sound, responsible decisions. In our previous work [1], [2], we translated Jeremy Bentham's utilitarianism into a decisionmaking algorithm, thereby operationalizing this ethical framework into an executable model for artificial agents. We demonstrated the viability of this *algorithmifying* of ethical theories into algorithms and further went on to translate ethical altruism (pure selflessness) and ethical egoism (pure selfishness) into algorithms. We examined how different ethical strategies impact both individual agent decisions and broader system outcomes.

Our experimentation indicated that moderate levels of greed yield more favorable societal outcomes than the extremes of pure altruism or pure egoism. A balance between selflessness and selfishness (embodied by utilitarianism) provided the best systemwide outcomes in our digital terrarium: the Sugarscape agent-based model which simulates an artificial society. Sugarscape provides a computational environment within which many thousands of artificial agents interact. Individual agent behaviors lead to emergent societal properties.

Our previous work is limited by agents having a static level of greediness throughout their lifetimes. We build upon previous work by shifting from a static to a dynamic treatment of agent greed. In real-world environments, conditions are constantly changing, and the ideal balance between greed and cooperation is often context-dependent. An agent must therefore adapt their greed to surrounding circumstances. For example, during heavy traffic, a self-driving car can reduce its greediness by choosing a longer route or adjusting its speed. This contributes to a smoother traffic flow and thereby improves travel efficiency for both itself and others. Individually cooperative acts create systemwide benefits.

We extend our ethical algorithms by endowing each agent with an initial predisposition toward greed which shifts in response to changing circumstances. Each agent can adapt its selfishness over time, allowing it to adjust its strategy in response to environmental feedback. This mechanism benefits both the agent and the system as a whole. Through a shock-

ingly simple adaptive mechanism, agents assess whether their current operational status requires more greedy behavior or a shift toward cooperation. This dynamic mechanism more accurately reflects the fluctuating balance between self-interest and cooperation observed in the real world.

We use the Sugarscape agent-based model to assess whether dynamic greed promotes more stable and successful multi-agent systems. We consider a variety of societal-scale metrics for success across the three ethical algorithms (utilitarian, altruist, and egoist) with both static and dynamic greediness for each. Compared to static agents, those capable of adapting their behavior demonstrate greater outcomes achieving up to 31.5% higher survival rates, supporting populations up to $2\times$ larger, living $1.7\times$ longer, and are $1.6\times$ happier. These results suggest that modulating self-interest over time helps avoid the downsides of rigid adherence to a single strategy, leading to more resilient systems. We conclude with observations about multi-agent AI systems beyond the Sugarscape agent-based model we use for our experimentation, demonstrating the broader applicability of our adaptive greediness approach.

II. Related Work

Greed has been studied across a number of disciplines such as business [3], biology [4], [5], philosophy [6], and psychology [7]. In evolutionary biology, the Green-Beard Effect [8] explores the development of altruism via selection. Interdisciplinary thought experiments like the Prisoner's Dilemma [9], [10] combine observations on greed from game theory, philosophy, economics, and political science. Practically all fields have something to say about greed, whether in humans, other animals, or even autonomous machines.

We investigate greed from a computational perspective using the Sugarscape agent-based model. This provides a digital terrarium for thousands of agents to interact with each other in a shared computational environment. Sugarscape was originally introduced in *Growing Artificial Societies* [11]. It is useful for simulating and studying social phenomena at large-scale, particularly economic phenomena [12], [13], [14]. However, it has also been applied to the study of machine ethics [15]. One such problem unaddressed by other studies using Sugarscape is an exploration of agent greed (or self-preference) in multi-agent AI systems. Sugarscape is an ideal sandbox for studying the feasibility of multi-agent systems where agents may have conflicting missions and decisionmaking procedures. The simulated environment is complex enough to provide non-trivial ethical dilemmas for agents while being computationally tractable.

We adapted and expanded the original Sugarscape model using modern software engineering best practices [16]. From there, we included explicit ethical reasoning into Sugarscape's agents [1]. We presented an initial study into agent greed in previous work [2] which we now expand.

The sheer quantity of socio-ethical concerns surrounding the usage and deployment of AI systems demonstrates the necessity to ensure AI agents act ethically. Blackman provides a thorough, introductory treatment of many topics in the ethics of artificial intelligence [17]. One concern which has captured the public imagination through science-fiction [18] is that AI will become an extinction-level threat (the validity of which has been critiqued [19]). A more grounded concern is the tug-of-war between deliberate, cautious, cooperative innovation in AI versus fast-paced, competitive market solutions. Sustainable, ethical advancement requires the former approach [20].

To avoid catastrophic outcomes in AI requires diligent effort to ensure agents engage in moral reasoning. This is the remit of machine ethics. In this field, one might ask when it is acceptable for an agent to lie to a human [21]. Do we find it acceptable for an AI chatbot to engage in little white lies like many humans do?

AI tools offer collaborative power to explore alternative ways of structuring socio-political systems [22]. However, even with humans in the loop, these tools may also reinforce already existing power imbalances. They do not provide unbiased, panacea solutions.

Machine ethics is filled with other similarly thorny issues, many of which are covered by Anderson and Anderson [23]. To advance the goal of machine ethics, we translated the ethical theory of utilitarianism into a decisionmaking algorithm for Sugarscape agents [1], [2] to demonstrate the viability of *algorithmifying* human ethical reasoning. We expand upon this previous work and present a more complete consideration of agent greed in this utilitarian framework. We introduce agents which adapt their selfishness in response to their surroundings while also adhering to a moral framework. This allows us to more fully replicate and expand upon the literature on greed across disciplines and apply these lessons learned to multi-agent AI systems.

III. Sugarscape

The Sugarscape agent-based model simulates an artificial society where thousands of agents interact with their environment and each other. These (simple) interactions lead to complex, emergent behaviors in the society. The simulation environment is represented by a two-dimensional $n \times m$ toroidal grid. Spread across the environment's cells are two resources: *sugar* and *spice*. Each cell is provided an initial amount of both resources (which may be zero). Cells regenerate their resources over time up to their original amount. Depending on the user-defined placement of resources, the environment may be especially hostile or hospitable to life.

A Sugarscape simulation run begins with a number of starting agents being randomly assigned to cells. Each agent aims to live as long as possible following simple rule-based mechanisms to accomplish this mission. In a default configuration, agents strive to survive by acting greedily.

Agents are configured with a *metabolism* for both sugar and spice. They eat some of the sugar and spice in their stockpile according to their metabolisms each simulation timestep. After this, they move to a new cell in order to gather more resources. They can move to a new cell according to their *vision* and *movement* which dictate how far they can see and travel, respectively. After moving, agents engage in a number of other

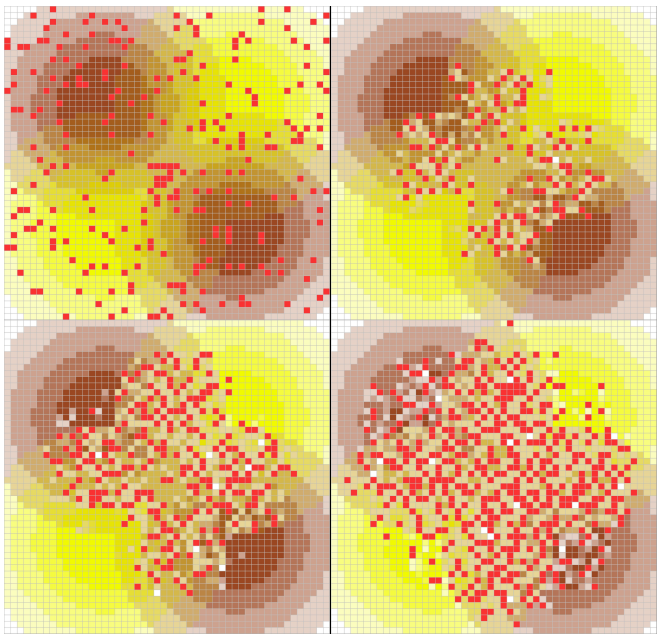

Fig. 1. A Sugarscape Society from Start to Finish

behaviors (if these features are enabled). These include: reproducing, making friends, spreading diseases, trading, lending, exerting cultural pressure, and engaging in combat. Previous work [16] details these features.

Figure 1 shows snapshots of timesteps across a brief simulation run. The top left displays the initial configuration of agents (red), sugar (yellow), spice (brown), and cells with both resources (tan). A greater saturation of color represents a larger number of resources at a cell. The top right shows the simulation 50 timesteps in where the population has coalesced near the resource-rich center but has also sustained casualties due to selection, starvation, combat, and disease. The bottom left presents the artificial society at 125 timesteps where it is rebounding from the initial die-off. The bottom right shows the end of the simulation at timestep 200 where a self-sustaining society has been established. In a greedy society, this happy outcome is far from guaranteed [2].

IV. Agent Decision Models

In previous work [1], we added a layer in the software architecture of Sugarscape agents to implement *decision models*. A decision model dictates an agent's behavior when they perform an action. The default behavior from the Sugarscape engine is simple, static, and greedy. Agents select their actions according to a calculation of how many resources they will gain. A decision model acts as a drop-in replacement for this behavior. Our first decision model was based on utilitarianism as presented by English philosopher Jeremy Bentham [24].

A. Utilitarianism

When considering an action, utilitarian decisionmaking focuses on the likely consequences of that action. Particularly, a utilitarian will determine the consequences borne by everyone impacted by the action. Bentham believed one could quantify these consequences, assigning positive values to happy outcomes and negative values to unhappy ones. These values are tabulated to find a composite score called the *utility* of the action. An ethical act according to a utilitarian viewpoint is one which results in more happiness than unhappiness. Bentham provides a pseudo-algorithm for calculating utility [24] which is called the *hedonic calculus*.[1]

Algorithm 1 shows the decisionmaking procedure for a utilitarian agent. The decisionmaker considers all affected by the potential action and tallies up the resulting (un)happiness likely to occur. The algorithm allows for the decisionmaker to apply their greed to the utility calculation, but true utilitarians are morally compelled to treat all actors equally (even themselves). So, a utilitarian would have a *selfishness factor* of 0.5 meaning all agents are treated equally.

Algorithm 1 Bentham's Hedonic Calculus

Given an action a drawn from a set A of actions
Given a decisionmaking agent d
Given a set P of people most affected by action a
$utility \leftarrow 0$
$s \leftarrow d$'s selfishness factor
for all $p \in P$ **do**
    $h_{p,a} \leftarrow p$'s happiness resulting from action a
    if $p \neq d$ **then**
        $utility \leftarrow utility + ((1 - s) * h_{p,a})$
    else
        $utility \leftarrow utility + (s * h_{p,a})$
    end if
end for
return $utility$

B. Ethical Altruism

Our second implemented decision model is ethical altruism. Unlike utilitarians, altruists care only for the consequences of others. They eschew any consequences (positive or negative) for themselves. They are entirely selfless agents. As such, their decisionmaking procedure is a simple modification from the utilitarian approach. Following the same procedure as Algorithm 1, they have a 0.0 selfishness factor.

C. Ethical Egoism

Our third decision model, ethical egoism, is likewise a simple twist on utilitarianism. An egoist has a 1.0 selfishness factor. Egoists care only for their own consequences and never for the consequences of others. They are emblematic of pure greediness, like the default Sugarscape behavior. But, they use the more sophisticated Algorithm 1 to determine their next move. Previous work has shown this more refined greedy approach is a better heuristic of optimal selfish actions [1].

[1] Also called the *felicific calculus* or the *utility calculus* in the literature.

V. Dynamic Selfishness

Each of the three decision models is static; an agent's selfishness never changes. This inflexibility can mean that an agent may starve to let another live. While this adheres to the core principles of utilitarian and altruist approaches, this rigid adherence is not always borne in reality. Many people prioritize their own needs when in desperate situations.

They may rationalize in this dire moment that temporary greediness is ethical, and many would concur! Tolerance is commonplace for violent self-defense or traffic violations in a medical emergency. This is exemplified in popular fiction by the character Jean Valjean in Victor Hugo's *Les Misérables* [25]. He steals bread in order to feed his starving family. He is imprisoned and continually persecuted for this act of desperation, but Valjean shows this single action does not define him nor condemn him to perpetual immorality.

We expand our consideration of artificial agent greed by allowing agents to adapt their selfishness according to their circumstances. The source material for Sugarscape [11] demonstrates that the model leads to rich, emergent societal behaviors from simple agent rules. We follow this methodology by making a simple change to the agent decision models.

We allow for an agent to evaluate its circumstances at the end of its turn each timestep by considering the change in its time to live (TTL). An agent's TTL represents the number of timesteps it can continue to survive without gaining any new resources. Since the agent metabolizes resources each timestep, a change in TTL indicates whether an agent is becoming closer or further away from starvation. Sugar and spice are also economic currencies, so TTL is also a barometer for an agent's gain or loss of wealth.

Algorithm 2 Dynamic Agent Selfishness Factor

Given a dynamic agent a
Given a's selfishness factor s
Given a's current time to live ttl_{curr}
Given a's last calculated time to live ttl_{last}
if $ttl_{curr} < ttl_{last}$ **and** $s < 1.0$ **then**
    $s \leftarrow s + 0.01$
else if $ttl_{curr} \geq ttl_{last}$ **and** $s > 0.0$ **then**
    $s \leftarrow s - 0.01$
end if
$ttl_{last} \leftarrow ttl_{curr}$

Algorithm 2 presents the simple modification to the decision models. This occurs at the end of an agent's turn. Each agent compares their TTL for the current turn to their TTL from last turn. If it decreased, that means they did not collect enough resources to maintain their same level of survival. In response, they choose to become slightly more selfish their next turn.

Likewise, if their TTL is the same or increases, they have made their position better. There may be other agents who did not get the resources they need. The agent chooses to decrease their selfishness in case others need to claim more resources next turn. This mechanism allows extra agent dynamicity which was not possible before.

VI. Evaluation

To evaluate the effectiveness of dynamic agent greed in Sugarscape, we experiment across 200 seeds. Each seed represents an initial starting setup of the simulation environment. Our implementation of Sugarscape is deterministic, so differences between simulation runs for the same seed are truly indicative of divergent agent choices. For each seed, we ran 6 different agent decision models:

1) Static altruist agents (immutable 0.0 selfishness)
2) Dynamic altruist agents (starting at 0.0 selfishness)
3) Static utilitarian agents (immutable 0.5 selfishness)
4) Dynamic utilitarian agents (starting at 0.5 selfishness)
5) Static egoist agents (immutable 1.0 selfishness)
6) Dynamic egoist agents (starting at 1.0 selfishness)

Each simulation ran for 5,000 timesteps and began with 250 initial agents. Agents are configured to live up to 100 timesteps, so this ensures at least 50 complete generations of agents in the simulation. The majority of other Sugarscape features are enabled such as: reproduction, trading, lending, combat, tribes, and disease. Combat, tribes, and disease are sources of hardship in an agent's life and demonstrate the need for cooperation. A detailed explanation of these features can be found in previous work [16].

Our experiments test whether agents with the ability to adapt their selfishness to their circumstances can overcome the failures of societies where agents are fixed in their ways. In previous work [1], [2], we demonstrated that entirely altruistic and entirely egoistic societies are prone to failure. Agents are either recklessly selfless or overly aggressive in these societies, respectively. As such, they collapse far more frequently than in societies where agents cooperate in a utilitarian manner using Algorithm 1 to calculate utility for every agent involved.

VII. Results

We evaluate societal success across a number of commonsense metrics. While these are not exhaustive, they demonstrate the usefulness of dynamic agent greed. These results first reproduce those found in previous work [2], then build upon the observation that utilitarian cooperation produced the most successful Sugarscape societies, and finally serve as a touchstone to similar lines of research in other disciplines.

TABLE I
Societal Survival

Behavior	Selfishness	Dynamic?	Extinct	Worse	Better
Altruist	0.0	–	83	10	107
Altruist	0.0	✓	61	1	138
Utilitarian	0.5	–	0	0	200
Utilitarian	0.5	✓	13	0	187
Egoist	1.0	–	89	1	110
Egoist	1.0	✓	26	1	173

A. Societal Survival

A critical metric for gauging societal success is whether or not a society made it to the end of the simulation. Table I

shows the survival rates for each experimental configuration sorted into three buckets: extinct, worse, and better. Extinct societies did not make it to the end of the simulation. Worse societies end the simulation with fewer than (or equal to) the number of starting agents. Better societies end with more agents than at the start.

We reproduce the finding from previous work [1], [2] showing that static utilitarian societies (with a selfishness of 0.5) are the benchmark for success. With perfect cooperation, agents are able to overcome a particularly tumultuous simulation start.

The beginning of every simulation was prone to death from combat, starvation, and selection. Since agents are scattered randomly at the start, some may begin life in a barren wasteland while other lucky ones are born in lands of plenty. We term this the *murderous period*. The static altruist (0.0 selfishness) and static egoist (1.0 selfishness) societies also reproduce the findings from our previous work. They do not fare particularly well with many societies going extinct, often during the murderous period.

In the case of altruism, agents often give up on choices necessary for their survival if those choices even mildly inconvenience another agent. Suicide is a preferred alternative for the dogmatic altruist. In egoist societies, rampant competition prevents starving agents from getting resources they need. The population enters a tailspin toward extinction as the few remaining strong agents cannibalize each other at the end.

When society does overcome these initial obstacles in altruist and egoist societies, it is often at the expense of a large number of less fortunate agents. Many of these were (self-)sacrificed for the remainder to thrive. Further discussion of these quirks inherent to purely altruistic and purely egoistic behavior can be found in previous work [1].

The dynamic utilitarian (0.5 starting selfishness) societies perform slightly worse than their static counterparts. This is caused by a significant drop in selfishness during the murderous period which deviates from a more optimal strategy of balancing greed and cooperation at the critical beginning of the simulation. The other dynamic societies are capable of significant improvement upon their static counterparts. The dynamic egoist societies mostly avoid the disastrous outcomes of the static egoists. Dynamic altruists, while performing far better, do not come as close to the static utilitarian benchmark. We show this is due to agents not becoming selfish enough during the murderous period.

B. Dynamic Selfishness

Since the dynamic variants of each configuration varied from the static counterparts in societal survival, it stands to reason that the mean selfishness in society changed for these. Figure 2 shows the mean selfishness factor per timestep for all seeds per decision model. Societies which went extinct have a zero value propagated out to timestep 5,000 from when they went extinct. This mildly affects the magnitude of certain values but does not affect the ordering in the graphs. This methodology applies to all our graphs in order to prevent cherrypicking of results.

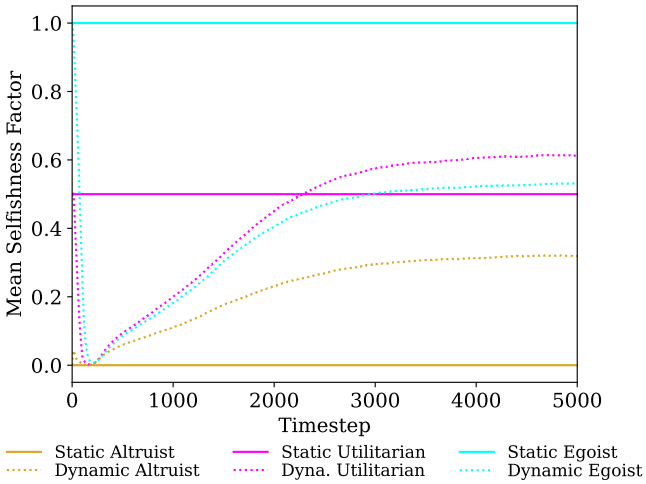

Fig. 2. Mean Agent Selfishness

For the dynamic utilitarian and dynamic egoist models, their selfishness nosedives during the beginning of the murderous period. The sharp decline in selfishness is caused by the harsh initial conditions of the simulation. During the murderous period, selection and predation leave relatively few agents who acquire many resources and increase their TTLs. For dynamic altruist societies, there was an slight bump in selfishness at the beginning of the simulation. All models increased in selfishness after the murderous period and reached a steady mean selfishness roughly halfway through the simulation.

For dynamic egoist societies, the steep decrease in selfishness during the murderous period prevents agents from engaging in conflict when society is at its most precarious. Eventually, these societies settle into a state quite similar to the static utilitarians at roughly 52% mean selfishness. Given the utilitarian model's proven record of success, the capability of the dynamic egoists to reach this seemingly ideal value indicates the success of the dynamic selfishness approach.

The dynamic altruists increasing their selfishness reproduces previous work [1], [2] which showed that at least a modicum of self-interest was necessary for societal survival and thriving. The small bump in selfishness at the beginning of the simulation demonstrates the importance of prioritizing survival in the murderous period. When the environment poses the harshest challenges, self-preservation becomes more important than complete deference to others. Once societies struggle through the murderous period, agents adopt a cooperative, slightly self-interested mean selfishness of roughly 25% which proves necessary for long-term success.

While the dynamic altruist model improves upon its static counterpart across all metrics, its survival rate and performance across metrics are still worse than the utilitarian models. In future experimentation, we may alter how quickly agents can change their selfishness factor. We suspect this will help altruists rapidly adapt to the murderous period then relax after the danger has passed.

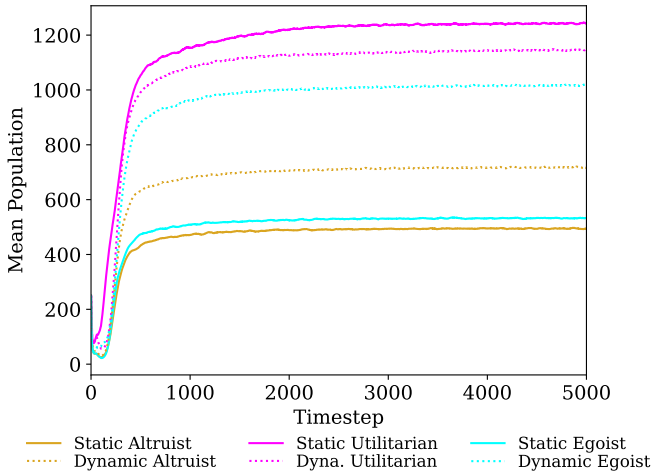

Fig. 3. Mean Societal Population

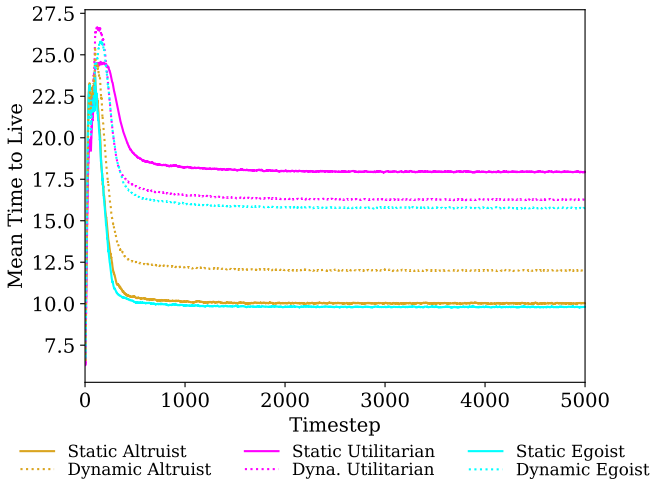

Fig. 4. Mean Agent Time to Live

C. Population

Dynamic selfishness does not only lead to increased societal survival; it can also yield larger societies. Figure 3 shows the mean population across all timesteps for all seeds per model. The static utilitarian model provides the benchmark of success at nearly 1,200 agents. Dynamic utilitarian societies nearly meet this performance, however the dynamic increase in greed causes these societies to drift from the purely egalitarian, static utilitarian results. Perfect balance between greediness and cooperation yields the largest population.

Dynamic egoism almost closes the gap between static egoism and utilitarianism due to the decrease in agent selfishness in the murderous period. Dropping to an average of 52% selfishness long-term allows the dynamic egoists to reap the benefits of cooperation. While dynamic altruism does not reach as much success, it improves upon the performance of both static altruism and static egoism. The brief bump in greediness during the murderous period encourages agents to gather more resources. These resources are necessary for reproduction, so the agents which survive the murderous period are more likely to be capable of propagating society.

D. Time to Live

While a larger population size is a strong indicator of societal success, it does not say much about the prosperity of the agents in society. Mean time to live (TTL) presents a clearer picture of agent success within society and is directly related to the mechanism for dynamic selfishness. TTL is the number of timesteps an agent can continue to survive without collecting any additional resources. A higher TTL represents an agent that has lower metabolisms for sugar and spice, a greater stockpile of those resources, or both. Higher mean TTL is an unqualified good in society.

Figure 4 shows the mean TTL across all timesteps for all seeds per model. Again, the static utilitarian model sets the high bar and is closely followed by the dynamic utilitarian model. The slight decrease in mean TTL for the dynamic utilitarian model reinforces that the perfect balance between greed and cooperation embodied by the static utilitarian model is closer to an optimal amount of greediness in society.

Dynamic egoist societies practically match the dynamic utilitarian performance. By expressing more cooperation, the dynamic egoists allow less fortunate agents a better chance to improve their circumstances. Rather than either taking the best cell or killing weaker agents, the dynamic egoists become cooperative enough to encourage societal flourishing.

The other two static models remain at the bottom for performance in TTL. The struggles to reach a stable society lead to short agent lifetimes. Agents are generally closer to starvation even well after the murderous period. Dynamic altruism improves upon these conditions though not nearly as well as the other dynamic models largely because agents do not reach a strong enough level of self-preservation after the murderous period. Though these societies may be more stable than their static counterparts, agents are not flourishing.

E. Happiness

The final metric we consider also measures agent prosperity. We measure happiness as a coarse-grained metric for overall agent satisfaction with their life circumstances. An agent's happiness is a composite score along five different axes: combat, family, health, social, and wealth. Each axis of happiness is represented as a real number in the range $[-1.0, 1.0]$ and are summed to create the composite score.

Combat (un)happiness represents remorse from killing, and an agent loses 1 happiness when they engage in combat in a given timestep. Family happiness increases the larger the agent's living family (children and mates). If a family member is sick, the agent loses family happiness. When a family member dies, there is a permanent decrease in agent happiness. Health happiness is gained when an agent is healthy but lost when they are sick. If an agent has many friends, their social happiness increases. If they have too few, their happiness decreases. Wealth happiness is derived from how much higher

an agent's wealth is above the global average wealth. If it is lower than the average, they lose happiness. Full details on agent happiness are presented in previous work [26].

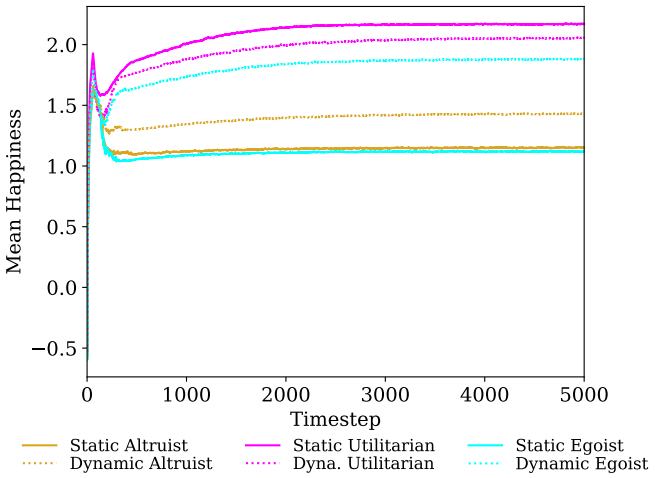

Fig. 5. Mean Agent Happiness

Figure 5 shows the mean agent happiness across all timesteps for all seeds per configuration. The maximum happiness is 4.0 in our experimental configuration.[2] The static utilitarian model did best with a mean score just above 2.0 happiness. Its dynamic counterpart nearly matches this value, and the dynamic egoist model closely follows. Like with mean TTL, this demonstrates that agents in these societies are generally prosperous. However, these two dynamic models slightly deviate from the optimal greedy-to-cooperative ratio of the static utilitarian model though they come close. This leads to slightly worse outcomes in mean happiness.

The dynamic altruist model does better than the other two static models. Yet, it does not reach the success of the other three models. It is worth noting, however, that all societies have at least some degree of mean happiness. Once the values are settled, no model results in a mean *un*happiness across society (though there may be individual unhappy agents).

VIII. Limitations

We have demonstrated that agents adapting their greediness to achieve their goals can lead to significant system-level benefits. In Sugarscape, dynamic altruistic and dynamic egoistic artificial societies experienced greater flourishing than their static counterparts. However, we also note there are limitations to our implementation of dynamic greed.

Investigating greed from a computational perspective is straightforward when the computational landscape allows only quantitative, economic decisionmaking. In Sugarscape, agents' actions are purely economic in nature, leading each to act as *Homo economicus* [27]. Changing an agent's behavior to follow an ethical theory (namely, utilitarianism) and to be dynamically greedy does not change the action set for the agent. Their choices remain economized: do that which provides the most resources (whether for yourself or for all).

Investigating greed is harder when applied to non-economic decisionmaking. If agents made choices based on happiness or making more friends or prioritizing reproduction, Sugarscape would become a richer model for future exploration. Each of these three different priorities reflects choices that are, at best, tangential to economic benefits. In the case of reproduction, which costs significant resources, it is economically disadvantageous yet socially necessary for the continued survival of society. In future work, we seek to introduce more varied, non-economic incentives for agent decisionmaking to increase the explanatory power of Sugarscape agents.

Our mechanism for varying greed is remarkably simple yet leads to pronounced results. By tracking an agent's time to live, they can react to times of plenty (become more selfless) and times of need (become more selfish). One could imagine a more complex mechanism could result in even better performance. It may also address the lackluster improvements by the dynamic altruist model. A more fine-grained approach could serve as a better heuristic toward optimal choices, and we note this as an avenue for future work.

Greed is regarded differently in other ethical theories. In the ethical models we presented, the selfishness of the agent simply places them at a different point along a spectrum. Altruism and egoism mark the extremes while utilitarianism lies in the middle. The underlying decisionmaking mechanism remains the same across this spectrum.

Kantian deontology, however, is wholly averse to greed; selfish acts treat others as mere means to one's own greedy ends. This flies in the face of Kant's *categorical imperative*. The ethics of Confucius places importance upon dyad relationships (such as parent and child or commander and soldier) and avoids greed. Selfishness harms structured relationships like a child's disobedience of a parent's request given out of care for the child's flourishing. In future work, we seek to translate other ethical theories into decisionmaking algorithms in Sugarscape. This is a thrilling proposition but will require us to reckon with radically different perspectives on how agents ought to interact with their environment and each other.

IX. Broader Applicability

Dynamic agent greed within an ethical framework has applications beyond Sugarscape. Multi-agent AI systems will be populated by agents designed by different teams and may have radically different missions. These differences will lead to changing circumstances within the computational environment and will affect interactions between agents. It may be preferable for agents to be able to react to these changes rather than simply stick to inflexible behaviors.

For instance, a smart energy grid is a smart environment where many autonomous agents interact. These include energy controllers across buildings, stationary smart devices, and mobile agents like self-driving vehicles. The shared environment (the grid) provides a limited resource: energy. Agents must balance their energy demands with the stability of the grid. Power

[2] Agents can be configured to gain happiness from killing but are not here.

delivery for humans is also a necessity. Dynamic selfishness allows these agents to use more energy to maximize their goals in normal circumstances but behave more cooperatively during peak demand or during energy shortages.

Self-driving cars must dynamically balance assertiveness and cooperation on the road. If an autonomous vehicle is overly cautious and deferential, it can lead to gridlock or being exploited by human drivers. Waymo's self-driving cars have sometimes been too polite, getting stuck behind obstacles or in a standoff of politeness where a human would normally nudge ahead [28]. Such selfless behavior even allows opportunistic humans to take advantage. These situations show why an adaptive approach (sometimes yielding, sometimes pushing forward) is needed. It is also a better facsimile of human behavior. By contextually adjusting their selfishness, autonomous vehicles can keep traffic flowing smoothly, avoid deadlocks, and avoid resorting to dangerous aggression.

Dynamic greed is not particularly suitable for AI systems which can be mapped to a zero-sum game. For instance, DeepMind's AlphaGo [29] has a mission to win at the board game Go. This zero-sum environment dictates any gain for an agent comes at an equal loss for the other. Dynamic greed is unnecessary. Optimality, from a game theoretic perspective, requires pure greed to win the game. This is especially true when the game is one-shot as opposed to iterated [30], [31]. Cooperating with the opponent, dialing back on greed, will only lead to the agent's loss.

X. Reproducibility of Experimentation

Our implementation of Sugarscape guarantees determinism meaning our results are reproducible.[3] All software requirements can be found in the `README` in the software repository. Our results were produced using the the `v2025.3` release. Newer versions will produce similar, but not exact, results.

We also provide all JavaScript Object Notation (JSON) configuration files to reproduce our experiments.[4] These configurations can be found in the `dynamic-greed.zip` archive. When running Sugarscape, it takes in one configuration file as input. For a single run of Sugarscape with a configuration, run the following (substituting the system's Python 3 alias):

```
> python sugarscape.py -c <CONFIG>
```

To reproduce our dataset, the repository configuration `config.json` needs to be modified. The `numParallelSimJobs` property defines the local parallelism allowed during data collection. Scale this value to the number of CPU cores or hardware threads available, being sure to leave enough to effectively run the operating system. Sugarscape is memory-efficient, so the bottleneck during data collection is available cores. Unzip all the configuration files in `dynamic-greed.zip` into the repository's `data` directory prior to execution.

[3]Software at: https://github.com/digital-terraria-lab/sugarscape.

[4]Configuration files at: https://github.com/digital-terraria-lab/datasets.

Run `make setup` to check for the system's Python 3 installation. This updates the `Makefile` and `config.json` accordingly. Run `make data` to do the data collection. Run `make plots` to generate graphs (requires `matplotlib` to complete). We note some modification to the plotting script in the `plots` directory is needed to reproduce exact formatting of our graphs. To reproduce the dataset and graphs, run:

```
> make setup
> make data
> make plots
```

Given that local data collection is likely to take a long time, we recommend extending the data collection process for distributed execution. For instance, previous work [1], [2], [26] used the Open Science Grid [32] and the HTCondor batch system [33] to enhance data collection. Distributed job creation and submission can be handled by writing a simple management script adapted to the batch system used. We encourage this approach to aid in reproducing our results.

XI. Conclusions and Future Work

We demonstrated that dynamic agent greed results in far better performance across a variety of metrics for agents which begin as pure altruists or pure egoists. Utilitarian agents are better off remaining purely utilitarian, ideally balancing greed and cooperation. For egoist societies, a 48% reduction in greed led to more stable societies which were $2\times$ larger, where agents lived $1.7\times$ longer, and where agents were $1.6\times$ happier. For altruist societies, an initial bump of around 5% selfishness during the murderous period and later increase to roughly 25% selfishness led to an 11% increase in societal survival with small increases in the metrics observed.

Our results reproduced and extended those from previous work [1], [2]. A modicum of selfishness is necessary for long-term societal success in multi-agent AI systems. Too much or too little selfishness yields unstable systems with high rates of collapse. An egalitarian balance between the two extremes leads to far better outcomes, but dynamic agent selfishness comes close to meeting these benefits. In real-world use cases, the dynamicity of agent ethical reasoning according to their surroundings may prove more valuable than strict adherence to a pre-programmed degree of self-interest.

We intend to continue extending our observations on greed from a computational perspective as it relates to the creation of explicit ethical decisionmaking by computer agents. One avenue for future work is to allow for heterogeneity in the starting selfishness in society. By allowing for a spread of pure altruists, pure egoists, and every percentage in between at the start of the simulation, selection dynamics come into play. It is harder to predict clean rankings of selfishness like in the models presented in this work. Further, it may be shown that adaptability is necessary for success in these heterogeneous societies whereas static societies may succeed or fail largely as a consequence of a lucky roll of the dice.

References

[1] N. Kremer-Herman, A. Gupta, and E. R. Severson, "Blueprints for machine ethics: A digital terrarium for socio-ethical artificial agent decisionmaking," *IEEE Access*, 2024.

[2] N. Kremer-Herman and A. Gupta, "The need for greed in artificial decisionmakers," in *International Symposium on Technology and Society*. IEEE, 2024.

[3] H. Mintzberg, R. Simons, and K. Basu, "Beyond selfishness," *MIT Sloan Management Review*, 2002.

[4] F. Heylighen, "Evolution, selfishness and cooperation," *Journal of Ideas*, vol. 2, no. 4, pp. 70–76, 1992.

[5] H. Rachlin, "Altruism and selfishness," *Behavioral and brain sciences*, vol. 25, no. 2, pp. 239–250, 2002.

[6] A. Rand, *The virtue of selfishness*. Penguin, 1964.

[7] L. R. Caporael, R. M. Dawes, J. M. Orbell, and A. J. Van de Kragt, "Selfishness examined: Cooperation in the absence of egoistic incentives," *Behavioral and Brain Sciences*, vol. 12, no. 4, pp. 683–699, 1989.

[8] R. Dawkins, *The selfish gene*. Oxford university press, 2016.

[9] M. Flood, M. Dresher, A. Tucker, and F. Device, "Prisoner's dilemma: game theory," *Experimental Economics*, vol. 54, p. 13, 1950.

[10] R. Axelrod, "Effective choice in the prisoner's dilemma," *Journal of conflict resolution*, vol. 24, no. 1, pp. 3–25, 1980.

[11] J. M. Epstein and R. Axtell, *Growing artificial societies: social science from the bottom up*. Brookings Institution Press, 1996.

[12] E. Serrano and K. Satoh, "An agent-based model for exploring pension law and social security policies," in *New Frontiers in Artificial Intelligence: JSAI-isAI International Workshops, JURISIN, AI-Biz, LENLS, Kansei-AI*. Springer, 2020, pp. 50–63.

[13] M. Oremland and R. Laubenbacher, "Using difference equations to find optimal tax structures on the sugarscape," *Journal of Economic Interaction and Coordination*, vol. 9, pp. 233–253, 2014.

[14] A. Rahman, S. Setayeshi, and M. S. Zafarghandi, "Wealth adjustment using a synergy between communication, cooperation, and one-fifth of wealth variables in an artificial society," *AI & society*, vol. 24, pp. 151–164, 2009.

[15] J. A. Lasquety-Reyes, "Towards computer simulations of virtue ethics," *Open Philosophy*, vol. 2, no. 1, pp. 399–413, 2019.

[16] N. Kremer-Herman and A. Gupta, "Replacing sugarscape: A comprehensive, expansive, and transparent reimplementation," in *International Conference on Simulation Tools and Techniques*. Springer, 2023, pp. 79–92.

[17] R. Blackman, *Ethical Machines: Your Concise Guide to Totally Unbiased, Transparent, and Respectful AI*. Harvard Business Press, 2022.

[18] K. Capek and J. Cejkova, *RUR and the Vision of Artificial Life*. MIT Press, 2024.

[19] D. Schlienger, "Ai is a trojan horse – a rationale," in *2024 IEEE International Symposium on Technology and Society (ISTAS)*, 2024, pp. 1–6.

[20] M. I. A. Ferreira, "The quest for an ai ethics: Between benevolence and greed," in *Producing Artificial Intelligent Systems: The Roles of Benchmarking, Standardisation and Certification*. Springer, 2024, pp. 155–167.

[21] V. Bakir, K. Bennet, B. Bland, A. Laffer, P. Li, and A. McStay, "When is deception ok? developing the ieee recommended practice for ethical considerations of emulated empathy in partner-based general-purpose artificial intelligence systems (ieee p7014.1)," in *2024 IEEE International Symposium on Technology and Society (ISTAS)*, 2024, pp. 1–6.

[22] A. Mertzani and J. Pitt, "Social implications of socially-guided machine learning for innovation support," in *2024 IEEE International Symposium on Technology and Society (ISTAS)*, 2024, pp. 1–8.

[23] M. Anderson and S. L. Anderson, *Machine ethics*. Cambridge University Press, 2011.

[24] J. Bentham, "An introduction to the principles of morals and legislation," *London: Athlone*, 1789.

[25] V. Hugo, *Les Misérables*. A. Lacroix, Verboeckhoven & Cie., 1862.

[26] M. Milkoswki, A. Gupta, and N. Kremer-Herman, "A little bit goes a long way: Modeling universal basic income for noncooperative artificial agents," in *International Symposium on Ethics in Engineering, Science, and Technology*. IEEE, 2025.

[27] D. A. Urbina and A. Ruiz-Villaverde, "A critical review of homo economicus from five approaches," *American Journal of Economics and Sociology*, vol. 78, no. 1, pp. 63–93, 2019.

[28] M. Bell, "What i learned from 130 hours in a waymo," Nov 2024. [Online]. Available: https://www.mattbell.us/what-i-learned-from-130-hours-in-a-waymo

[29] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[30] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[31] P. Danielson, *Artificial morality: Virtuous robots for virtual games*. Routledge, 1992.

[32] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Wurthwein *et al.*, "The open science grid," in *Journal of Physics: Conference Series*, vol. 78, 2007.

[33] D. Thain, T. Tannenbaum, and M. Livny, "Condor and the grid," *Grid computing: Making the global infrastructure a reality*, pp. 299–335, 2003.