A Case for Embedding Moral Reasoning in Artificial Agents Engaged in Conflict

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

Maria Milkowski
University of Notre Dame
Notre Dame, IN (USA)
mmilkows@nd.edu

***Abstract*—Artificially intelligent agents produced by various designers, with various directives, are likely to come into conflict when deployed in a shared computational environment. When conflict occurs, people dependent on computerized decisionmakers will be harmed. We present a decisionmaking algorithm for agents to engage in explicit ethical reasoning and avoid wasteful conflict. We elaborate upon previous work demonstrating that the benefits of ethical agent behavior can be approximated by a social safety net in the computational environment. Our inspiration is universal basic income (UBI).**

We use the Sugarscape agent-based model to show the limits of using UBI as a surrogate for ethical behavior in multi-agent systems. Agents exposed to a high degree of conflict must cooperate to avoid self-destructive outcomes like total system collapse or catastrophic mission failure. Sugarscape societies where agents engage in ethical reasoning result in up to 4× greater system survival with half the conflict of societies where agents receive UBI but do not engage in an ethical reasoning. We conclude that digital safety nets like UBI are not helpful when agents are predisposed to conflict; cooperation is a necessity.

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

I. Introduction

The number of autonomous computer agents making decisions on behalf of humans has drastically increased in recent years. It is imperative that these artificially intelligent (AI) agents behave according to well-understood ethical principles (such as those from Kantian deontology or virtue ethics). By doing so, scholars of ethics can effectively reason about how an agent will behave when deployed into a new environment. This enables transparent, explainable artificial intelligence.

An AI agent is an autonomous, goal-oriented computer process, software object, or algorithm which interacts with an environment and may improve itself over time. The vocabulary in the artificial intelligence domain is changing rapidly, resulting in some disparate terminology. We liberally accept AI agents as displaying a broad range of intelligent behaviors and investigate how to imbue them with a capacity for moral reasoning. Our topics of discussion can apply to various kinds of AI agents across this broadly defined class.

We build upon previous work where we demonstrated a proof of concept for agents which engage in moral reasoning. We translated the principles of utilitarian ethics into a decisionmaking algorithm [1], [2]. This enabled agents in the Sugarscape agent-based model [3] to cooperate toward *the greatest happiness for the greatest number* as posited by philosopher Jeremy Bentham [4]. We showed it is viable for agents to follow ethical principles in an environment where agents can come into conflict trying to fulfill their missions.

Our *algorithmifying* of ethical decisionmaking required that all agents cooperate in order to secure society-wide benefits. In reality, enforcing ethical behavior at global scale is impractical. AI agents are designed by various organizations with different missions. Conflict occurs when agent priorities come into opposition in a shared computational environment.

Rather than enforcing globally ethical behavior, the environment can provide a safety net to address unethical choices made by some (or perhaps many) greedy agents. We take inspiration from universal basic income (UBI) to provide an environmental mechanism which could act as a surrogate for the benefits of ethical behavior [5]. Instead of enforcing ethical behavior, all agents are allowed to act greedily but receive UBI in the form of resources from the environment. In previous work, we introduced this mechanism in a computational environment with limited agent conflict [5].

We evaluate whether UBI can remain a useful stand-in for globally ethical agent behavior when agent conflict is increased. We provide agents with greater opportunity for, and rewards from, killing other agents. With metrics including societal survival, life expectancy, and death rate, we demonstrate that even mild increases in combat squanders the benefits of social policy interventions like UBI for agents. Ethical behavior from the agents is necessary for any semblance of societal-scale success.

We use Sugarscape to set loose thousands of agents and observe how they interact. We describe the ethical design of these agents and their respective macro-scale consequences. While simple relative to full-fledged AI agents deployed in real-world contexts, Sugarscape is complex enough to avoid trivializing the hard work of constructing agents which engage in explicit moral reasoning. Our observations provide an empirical, controlled glimpse into real concerns regarding agent conflict relevant across domains. As an extreme example, AI agents deployed in warfare necessarily come into conflict and should behave in ways which limit collateral loss of human lives. More mundane instances of AI agent conflict abound as AI tools are deployed on behalf of consumers and producers, artists and digital copycats, and so on.

Even when minimally incrementing two dials of agent conflict (the number of agent tribes and amount of loot gained from killing), Sugarscape societies with UBI survive up to $4\times$ less, result in a roughly 8% lower life expectancy, and engage in nearly $2\times$ as much combat compared to cooperative, utilitarian societies. We demonstrate the necessity for cooperative agents across a variety of metrics. We finally connect the conflict experienced by Sugarscape agents to AI agents broadly with an eye toward a universal adoption of explicit ethical reasoning in all deployed AI agents when making decisions on behalf of, and when interacting with, humans.

A. Working Definition of Universal Basic Income

Universal basic income is a social welfare policy that has become a popular source of discussion. In the United States, its recent popularization was invigorated by American businessman and politician Andrew Yang's *The War on Normal People* [6]. Yang suggests that UBI can offset labor devaluation caused by job automation and can increase the quality of life of the general population in a world where many see their careers as *bullshit jobs* [7]. We refer to Hasdell et al. [8] for a working definition of UBI as a social policy. They define UBI as having the five following qualities:

1) It is universal and does not target a specific population.
2) It is unconditional.
3) It is a cash payment, with no stipulations how it is spent.
4) It is paid on an individual basis.
5) It is a recurring payment.

II. Related Work

The modeling of social welfare and fiscal policies is critical for understanding their potential longterm effects. Often, these policies are modeled mathematically or using microsimulations [9], [10], [11]. However, agent-based models (ABMs) allow for finer-grained explorations of individual behavior which can lead to unexpected emergent behaviors in the simulation [12], [13]. Social policies such as police funding [14] and the relationship between innovation, wages, and unemployment [15] have been effectively studied using ABMs.

Sugarscape is an ABM for simulating artificial societies and was first introduced by Epstein and Axtell in *Growing Artificial Societies* [16]. Since its introduction, Sugarscape has been used to study a number of social phenomena. For instance, it has been applied to social welfare [17], tax structures [18], and wealth disparity [19].

We apply Sugarscape in a new direction: machine ethics. The field of machine ethics is concerned with theoretical and practical challenges to creating AI agents which engage in ethical reasoning [20]. There are many socio-ethical concerns with the usage and deployment of artificial intelligence systems and machine learning [21] including fears (and rebuttals) of AI as an extinction-level threat [22]. Questions in machine ethics, such as when it is acceptable for an AI agent to lie to a human [23], demonstrate why developing ethical AI is difficult yet imperative. In response, the design of AI agents which engage in moral reasoning has become a burgeoning field of inquiry with some authors providing best practices for agent design [24] and others providing approaches to the writing of ethical reasoning algorithms [25].

We were inspired to use Sugarscape for machine ethics research by Lasquety-Reyes' use of the model to translate a key portion of Aristotelian virtue ethics into a decisionmaking algorithm [26]. We first adapted Sugarscape from *Growing Artificial Societies* and expanded it using modern software engineering best practices [3]. Then, like Lasquety-Reyes, we translated the core of an ethical theory into an algorithm to guide Sugarscape agent decisionmaking. We implemented a utilitarian algorithm [1], [2] which resulted in significant benefits compared to the default, greedy behavior of agents in the simulation. Sugarscape remains relevant for studying social phenomena and is an ideal platform for exploring topics in the burdgeoning field of machine ethics.

We directly expand upon previous work where we introduced universal basic income in Sugarscape as a surrogate for ethical behavior [5]. In our body of work, we found incredible benefits for a large-scale system of agents behaving cooperatively (according to utilitarian ethical principles). However, we also understood that cooperation across all agents is infeasible in real-world contexts where AI agents come from various design teams and have different motivations while operating in a shared environment (e.g. the factory floor, the battlefield, or the Internet).

We introduced UBI in Sugarscape to determine whether a mechanism in the computational environment could yield the same benefits of all agents cooperating without actually requiring cooperation. We saw significant benefits to using UBI where all agents acted selfishly, but it was also clear that ethical agent behavior led to better outcomes across some crucial metrics. Full details of these previous findings and a preliminary discussion of limitations using UBI as a surrogate for ethical behavior can be found in previous work [5].

III. Sugarscape

The Sugarscape agent-based model simulates an artificial society and displays emergent societal behaviors arising from individual agent behaviors. The simulation takes place on a two-dimensional $n \times m$ toroidal grid environment. The environment has two resources placed upon it: *sugar* and *spice*. Cells in the grid have an initial allocation of both resources (which might be zero). Cells regenerate over time up to their initial allocation. The environment may be particularly hostile or hospitable given the options enabled in a user-defined configuration file.

The simulation begins with an initial group of agents placed in random cells across the environment. These agents have a single aim: to live as long as possible. By default, they accomplish this through wholly greedy means.

Every agent has a configured *metabolism* for sugar and for spice. Each simulation timestep, they consume the sugar and spice they have on hand according to their metabolisms. They also move to a new cell to gather more resources according to their *vision* and *movement* which dictate how far

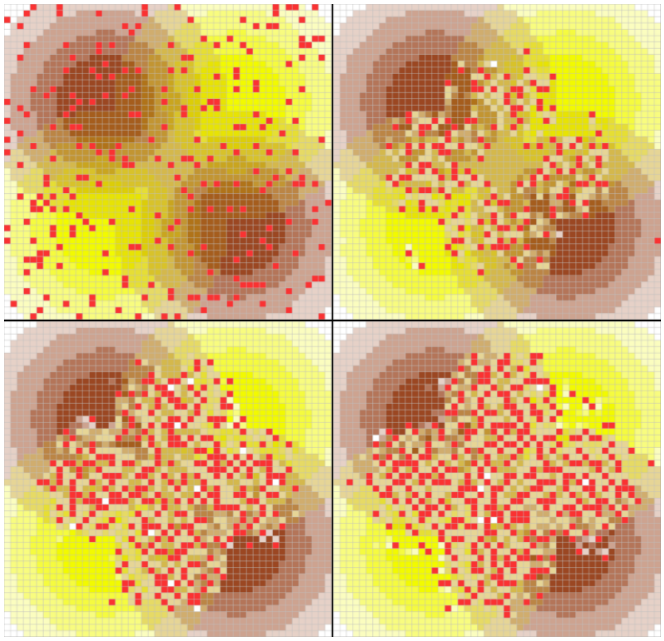

Fig. 1. A Sugarscape Society from Beginning to End

they can see and travel. Other behaviors are possible (when enabled) per timestep such as making friends, reproducing, communicating diseases, lending, trading, exerting cultural pressure, and engaging in combat.

Sugarscape provides a phenomenal sandbox for exploring agent-based behaviors, particularly since rich social dynamics arise from simple agent rules. The coarse granularity of the simulation allows for easily computable results while retaining strong metaphors to the real world [16]. A more complete discussion of the features in our Sugarscape implementation can be found in previous work [3].

Figure 1 shows snapshots of a complete simulation run. The top left shows the random proliferation of agents (red), sugar (yellow), spice (brown), and cells with both resources (tan). Greater saturation of color in a cell denotes a larger number of resources. The top right shows the society at timestep 50 where the population has shrunk due to combat, disease, and selection. Agents are clustering in resource-rich zones across the environment. The bottom left shows society at timestep 125 rebounding in the resource-rich zones. The bottom right shows the final position of the simulation at timestep 200 where a prosperous and successful society has taken root.

A. Universal Basic Income in Sugarscape

We used our working definition of UBI [8] when implementing it in our version of Sugarscape in previous work [5]. The simulation can be configured to provide UBI for sugar and spice (with differing amounts for both). The payment interval for both is also configurable. In our experimentation, we provide an equivalent amount of sugar and spice UBI and make payments every simulation timestep.

In previous work [5], we demonstrated that our UBI mechanism can be used as a surrogate for globally ethical behavior. Rather than relying upon every agent to act considerately toward their neighbors, the UBI policy instead provides a more achievable safety net for all agents even if some (or all) of their neighbors act greedily. In some metrics, artificial societies with UBI performed better than those with global cooperation. The previous work provides complete details which we eschew here for brevity [5].

B. Agent Conflict in Sugarscape

Conflict arises between agents through combat. During an agent's move, they calculate which cell will yield them the resources they need most for continued survival. If a cell has another agent on it, their accumulated wealth is added to the calculation for the cell. There is a limit to this added reward, which we call *combat loot*, representing how many added resources an attacking agent can steal while the remainder are lost. This limit is user-configurable before runtime.

An agent determines who it can attack based on its strength and the strength of a potential victim. Strength is coarsely determined by the agent's wealth. This is calculated by summing the agent's currently stored sugar and spice. So long as an attacker has more wealth than the defender, they will win. Combat outcomes are guaranteed, and an attacker will only attack if they will win.

One further aspect complicates combat: agent tribes. In combat, an agent can only attack another if they belong to a different tribe. There are a configurable number of social tribes in the simulation, and the initial spread of agents are sorted into these tribes such that there is an equal amount of agents in each at the start. An agent's tribe is determined by a bit vector representing its cultural preferences, called the agent's *tags*. These could be emblematic of musical tastes, political affiliation, religious beliefs, etc.

When an agent moves to a new cell, it can exert cultural influence upon its new neighbors. It does so by flipping a random bit in each neighbor's tags which may cause a neighbor to convert from their tribe to the agent's tribe if a threshold of bits match. Full details of this cultural transmission mechanism can be found in the Sugarscape source material [16].

C. Agent Cooperation in Sugarscape

By default, agents are greedy. They select moves according to maximizing their resources. In previous work [1], [2], we allowed for agents to engage in ethical reasoning.

We based this modification of agent behavior on the principles of utilitarianism as introduced by philosopher Jeremy Bentham [4]. He theorized the morality of an action can be quantified based purely on its consequences. Actions which produce happiness (called *utility*) for lots of people are deemed more ethical than those which produce very little happiness or those which produce more harm than good.

We translated Bentham's *hedonic calculus*, a pseudo-algorithm he crafted in the late 1700s, into a modern algorithm to guide agent behavior. With utilitarian decisionmaking enabled, agents choose their moves democratically. Algorithm 1 shows the procedure for calculating the utility for a potential

agent action. Full details of this algorithmic translation can be found in previous work [2].

Algorithm 1 Bentham's Hedonic Calculus

Given an action a from a set A of available actions
Given a decisionmaking agent d
Given a set P of people most affected by action a
$utility \leftarrow 0$
for all $p \in P$ **do**
    $h_{p,a} \leftarrow p$'s happiness resulting from action a
    $utility \leftarrow utility + h_{p,a}$
end for
return $utility$

IV. Evaluation

We experiment with the maximum combat loot and the number of tribes in Sugarscape as dials to ramp up agent conflict. Increased combat loot rewards conflict, making it more likely to occur. An increased number of tribes provides a greater number of targets for conflict as agents can only attack those from tribes other than their own.

By increasing conflict, we demonstrate the limits of UBI as a surrogate for ethical agent behavior. Additionally, we make a case that AI agents must engage in their own moral reasoning in particularly hostile conditions. Cooperation is a necessity for systemwide success in these situations.

We evaluate the limits of UBI and cooperative behavior across 200 random seeds. These seeds dictate the initial starting conditions of a simulation run and ensure that each run is deterministic. Differences between results in the same seed are truly indicative of different agent choices. Each simulation starts with 250 agents as in Figure 1 and ran for 5,000 timesteps. This ensures 50 generations of agents in the artificial society. We enabled most features in the simulation, including a variety of agent behaviors such as: combat, diseases, lending, reproduction, trade, and tribes. See previous work [3] for more details of these features.

For each seed, we experimented with 2 and 3 maximum combat loot as well as with 3 and 4 tribes in the starting population. In our experimentation, we found that simply increasing loot and the number of tribes by one was enough to dramatically affect societal outcomes. Ratcheting up conflict with further loot or tribes leads to scenarios where only the cooperative approach yields any societal success which is uninteresting for broader discussion.

We ran these experiments with agents receiving 1 sugar and 1 spice per timestep as a UBI payment. We compare these results to the control case where agents behave greedily without any UBI payment. Additionally, we provide the results for cooperative behavior using the utilitarian algorithm from previous work [2] as presented in Algorithm 1. The cooperative approach does not require UBI for societal success, so these societies never receive any.

V. Results

We consider societal success across a variety of metrics. While these are not exhaustive, they are indicative that there is a limit to how much conflict can be overcome by using UBI as a surrogate for ethical, cooperative behavior in society. After only minimal increases in the amount of conflict generated in society, the cooperative approach yields consistently strong outcomes while those of UBI rapidly diminish.

TABLE I
Societal Survival with Varying Tribes and Combat Loot

Behavior	UBI	Tribes	Loot	Extinct	Worse	Better
Greedy	0	3	2	71	2	127
Greedy	1	3	2	4	0	196
Cooperative	0	3	2	0	0	200
Greedy	0	4	2	103	3	94
Greedy	1	4	2	25	0	175
Cooperative	0	4	2	0	0	200
Greedy	0	3	3	157	3	40
Greedy	1	3	3	117	1	82
Cooperative	0	3	3	0	0	200
Greedy	0	4	3	177	1	22
Greedy	1	4	3	151	0	49
Cooperative	0	4	3	0	0	200

A. Societal Survival

Whether or not a society survives to the end of the simulation is a straightforward metric of societal success. We group configurations into three buckets based on survival: extinct, worse, or better. Extinct societies did not make it to timestep 5,000. Worse societies ended the simulation with a lower population than the start. Better societies ended the simulation with more agents than at the beginning.

Table I shows societal survival rates with a varying number of tribes and combat loot. The first block of results reproduces the finding of previous work [5] and represents the control case for the default greedy agent behavior, the control for the greedy behavior with a UBI payment, and the control for cooperative agent behavior without UBI. In previous work, we found that UBI acts as a remarkable surrogate for globally cooperative behavior while also being eminently more feasible in real-world use cases. Further, there were a number of other metrics in which societies with UBI performed better than societies where all agents simply cooperated.

Once incentives for combat are dialed up, the usefulness of UBI as a surrogate for ethical behavior plummets. The second block of results in Table I show a marked decrease in societal survival when the number of tribes is increased yet the combat loot remains the same as the control. By providing more targets for combat, the amount of combat naturally increases. For societies where agents act greedily, the additional conflict opportunity proves too much for many to bear. UBI payments mollify the negative impacts significantly, however there is still a decline in societal survival.

The third block of results in Table I show a return to three tribes as in the control case but a dialed up combat loot. Simply incrementing the amount of added resources an agent

can take as a reward for conflict proved enough to throw both greedy configurations into tailspins. Far more societies went extinct in this experiment showing that the expected reward of combat is the stronger driving factor for increasing violence in Sugarscape. These two greedy configurations perform so poorly that the cooperative approach in the face of such aggression is the only viable path toward societal success.

The fourth block of results in Table I combines the previous two experiments: the number of tribes is larger and the combat loot is incremented from the control. Combining these two effects practically halves the results of the third block for the two greedy approaches while the cooperative approach remains unassailed. The vicious worlds created by these configurations could only realistically be overcome by cooperation. In the few cases where societies survived in the two greedy configurations, the populations were relatively small with sharp wealth disparity.

Societal survival is a key metric for success. In artificial worlds lacking cooperation, survival is a roll of the dice where the odds are not in society's favor. Benefits like increased population size, life expectancy, or agent happiness are meaningless if there are no agents to enjoy them.

For the remainder of our results, we focus on the second block of societies in Table I. This is a sweet spot of increased conflict and a fair chance at survival for the two greedy configurations. The third and fourth blocks of societies demonstrate that cooperation is necessary for survival in high-conflict societies. They do not provide further useful data from the greedy approaches.

B. Life

Like societal survival, population size is a straightforward metric to measure success. Figure 2 shows the mean population size across all 200 seeds for each behavior and UBI configuration across all 5,000 timesteps the simulation ran. Our figures present the results of the second block in Table I with 4 tribes and 2 combat loot. In our figures, we label the greedy behavior with no UBI as *Greedy*, greedy behavior with UBI as *UBI*, and the utilitarian behavior as *Cooperative*. The greedy behavior without UBI configuration is our control case.

For extinct societies, a value of zero is propagated through from the timestep they went extinct to the end of the simulation. This impacts the degree of the charted values but does not change their ordering and counteracts cherrypicking only the most flattering results for the greedy configurations. This methodology is repeated in the remainder of our results.

We note that the beginning of the simulation leads to a population dip for all the configurations. We term this roughly 500 timestep stretch the *murderous period*. The beginning agents in society are scattered at random locations, with few starting resources, some carrying diseases, and the tribes are mixed together. These conditions are ripe for selection, starvation, and conflict. Those who survive the murderous period are responsible for whether society rebounds or collapses. Many societies which survive the murderous period make it through to the end of the simulation.

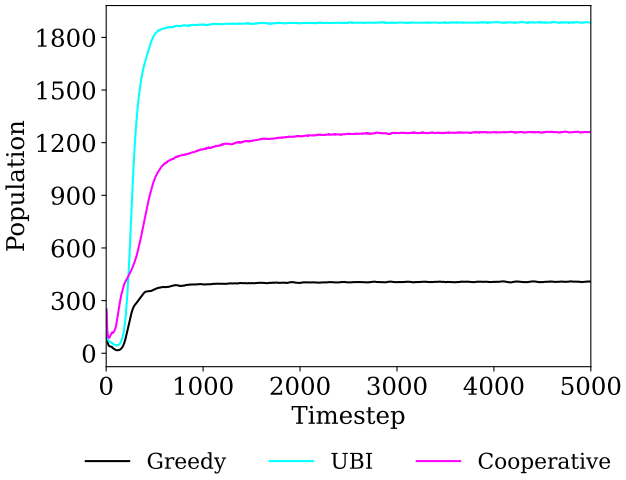

Fig. 2. Mean Societal Population

As discussed in previous work [5], providing a little bit of UBI allows agents to reproduce more easily over greedy societies without UBI. The resource barrier to reproduction is reduced, so we see a higher mean population than the other two configurations. While the societal survival is significantly lower with UBI compared to cooperation, perhaps the benefit of prodigiously populous societies outweighs the greater likelihood that society collapses in the murderous period.

The act of cooperating likewise addresses the difficulties of reproduction by allowing agents to more equitably gather resources. This leads to more mating behavior overall, resulting in a significant improvement upon the control (though not as high as the free resources provided by UBI which directly address the difficulty of reproduction). Without the aid of UBI or cooperative behavior, the greedy control configuration yields small societies which perform poorly across most metrics.

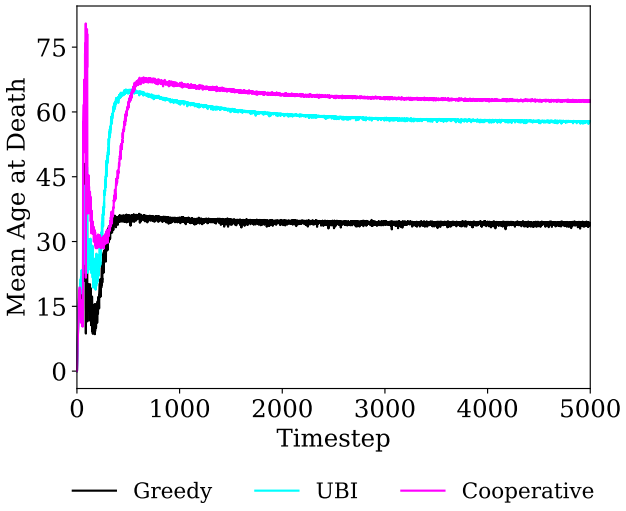

Fig. 3. Mean Agent Age at Death

C. Death

Large populations are not necessarily healthy ones. Life expectancy is another useful metric to gauge societal success. Figure 3 shows the mean age at death for agents across all seeds and timesteps for each behavior and UBI configuration.

The cooperative approach narrowly surpasses the UBI configuration in life expectancy, but both do quite well compared to the control case. The control represents the Hobbesian state of nature in greedy societies; agents lead lives that are *nasty, brutish, and short* [27]. The other two configurations practically double the life expectancy. UBI payments directly address any resource shortfalls agents experience, and the payments can feed agents at risk of starvation. Cooperation leads to a world where agents are less likely to engage in combat and can democratically decide that agents in need reach the resource-rich zones in the environment.

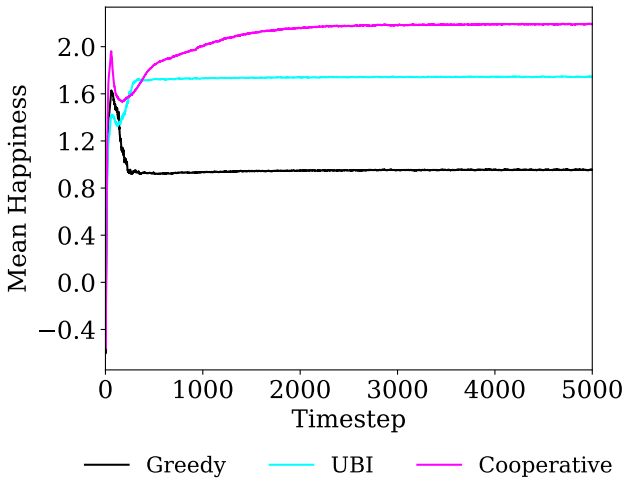

Fig. 4. Mean Agent Happiness

D. Happiness

To gain a more holistic understanding of agent outcomes, we measure their happiness. Agent happiness is measured along five axes to form a composite score: conflict, family, health, social, and wealth. Engaging in conflict results in remorse by the killing agent and decreases happiness. Having a large, living family improves happiness while having sick or deceased family members causes unhappiness. A healthy agent gains happiness while a sick agent loses it. Agents with lots of friends gain social happiness while those with few friends lose it. Finally, agents with wealth higher than the mean global wealth gain happiness while those who are poorer than the mean lose it. These measures for agent happiness are coarse-grained, and full details on their calculation and viability are found in previous work [5].

Agents who engage in combat or whose family members are victims of combat are likely to be unhappy. Conflict has a strong, negative impact on happiness. Figure 4 shows the mean agent happiness across all seeds and timesteps for each behavior and UBI configuration. The highest possible happiness is 4.0 for any agent.

The greedy configuration without UBI settles at a mean happiness less than a quarter the maximum possible score. While they avoid net unhappiness (a negative mean score), it is clear that happiness is depressed in these societies. The greedy approach with UBI does much better. The extra resources provided by the social safety net address many issues in an agent's life, like being able to reproduce more as shown in Figure 2. A larger family provides more happiness.

However, neither configuration matches the cooperative configuration. While not shown in this paper, cooperation yields lower wealth disparity than in the UBI configuration meaning a greater mean wealth happiness. There is also less conflict in these societies, leading to smaller penalties brought about by combat and moderate benefits to family and social happiness since family members are less likely to be killed and friends are more readily found.

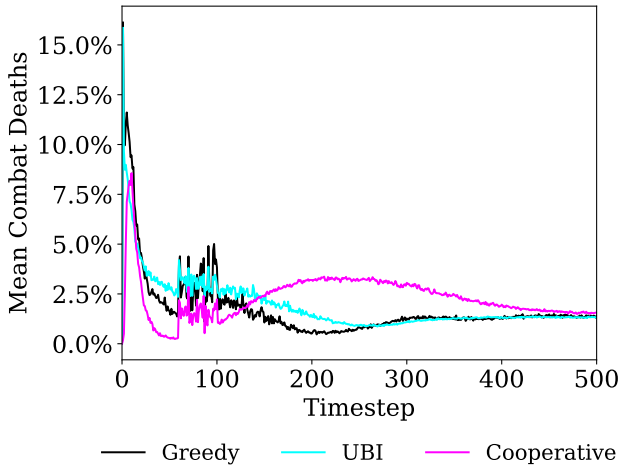

Fig. 5. Mean Agent Combat Deaths (% of Population)

E. Conflict

To demonstrate that cooperation results in less combat, we peek into the murderous period for answers. Figure 5 shows the mean combat deaths per timestep for all seeds per configuration for only the first 500 timesteps. The number of combat deaths is presented as a percentage of the population since the degree would be imbalanced as shown in Figure 2.

At the beginning of the simulation, the two greedy configurations have a spike of combat. Nearly 17% of the population is culled for each of the first 25 timesteps. The cooperative approach has half the conflict during this same time period.

The proliferation of combat begins to settle down in the two greedy configurations after 50 timesteps and approaches zero at the end of the murderous period. This is often due to one tribe reigning supreme and exterminating the others, resulting in no further combat being possible. In other cases, the lack of combat is due to the sheer reduction in population making agents less likely to encounter another viable victim.

The cooperative approach has a short-lived bump in combat in the middle of the murderous period after agents have settled. This is often due to more tribes surviving later into the simulation, making combat a viable agent action for longer. However, the degree of conflict is much smaller overall than in the greedy configurations. Though it lasts slightly longer before shrinking to almost zero, it is never as rampant nor as destructive as the greedy cases.

The key difference is that cooperative agents democratically select what action each of their neighbors should take. This allows for potential victims of combat to cry foul and for struggling agents to advocate for their own wellbeing, which is not possible in the greedy configurations. The cooperative approach does not eliminate the option of conflict nor does it manipulate the rewards from participating.

We can now claim there is truly less conflict in the cooperative configuration where agents adhere to utilitarian principles. UBI does not act as a valid surrogate for ethical behavior with regard to this metric. We note that substituting other principles in the cooperative approach, such as those inspired by Aristotelian virtue ethics [28], may yield different outcomes. We leave this exploration, and its comparison to the utilitarian approach, for future work.

VI. Limitations

By increasing conflict, we demonstrated that UBI is not an effective surrogate for ethical behavior when conflict is even mildly increased. However, we note that there are only two direct means of adjusting opportunities for conflict in Sugarscape: increasing combat loot and increasing the number of tribes. Increasing combat loot provides a greater incentive for agents to engage in conflict. Increasing the number of tribes provides agents with more targets for combat as no agent can kill a member of their own tribe. Both of these options affect Sugarscape's only method of direct conflict: combat.

The victor of combat is determined solely by their wealth. This reproduces social phenomena where those with the most resources at their disposal are relatively untouchable. However, there is a risk that the simulation may systemically advantage those agents who are born wealthy or born with the traits conducive to resource acquisition (i.e. high vision, high movement, and low metabolism).

Dialing up the combat loot and the number of tribes did not appreciably impact the amount of killing by the wealthy relative to the poor. In future work, we may change the combat strength calculation to avoid conflating variables relating to resources and conflict success. While having more resources often yields greater chance of conflict success (one need only look at modern military technology), in reality wealth does not provide a guarantee of victory as in Sugarscape.

One could interpret the cultural pressure agents exert upon each other as conflict. When moving, an agent tries to convert their new neighbors to their tribe. Other interactions could be interpreted as forms of non-violent conflict such as power imbalances between lenders and borrowers. In future work, we may explore adding a greater variety of direct conflict options for Sugarscape agents to bolster its explanatory power.

Sugarscape agents have a single goal: acquiring resources to survive. As such, their motivations are entirely economic. If agents could make decisions based on non-economic factors, a richer number of (potential conflict) interactions develop. For instance, quarantining when sick is widely considered an ethical behavior, yet it can economically disadvantage an agent. We intend to expand the currently limited conception of agent happiness to drive non-economic decisionmaking. This will allow for Sugarscape agents to become far more complex and more thoroughly represent ethical dilemmas.

Our implementation of UBI in Sugarscape is unrealistic for human societies. There is no government structure among the agents meaning no taxation, no competing interests over a shared budget, and so on. In future work, we may implement the capacity for agents to form government structures. This would radically alter the action set for agents (including adding new conflict mechanisms).

However, our UBI mechanism's departure from human reality does not diminish its worth in our study of machine ethics. UBI as presented acts as an environmental mechanism to address systemic inequalities. These include helping agents who have high metabolisms, low vision, low movement, or who were unlucky enough to be born in a resource-poor area. When generalizing this to other multi-agent AI systems, our UBI implementation is simply one method for the environment to encourage agent flourishing which happens to be inspired by human social policy. Many other potential mechanisms exist.

VII. Broader Applicability

Our findings provide insight beyond Sugarscape. Deployed AI agents come from various manufacturers, with different missions, yet operate in the same environment. For instance, a road may be populated by self-driving vehicles from a number of different companies. The engineers behind each product implemented the self-driving features independent from other companies, yet they all share the road.

We also see systems with a number of AI agents interacting to accomplish individual aspects of a larger mission. For example, in an AI-assisted shopping platform, a chatbot may take input from a user, parse a command from the input, direct another agent to perform some analysis upon a dataset (perhaps checking inventory and estimating future demand for an item), which then directs another agent to perform an action (such as purchasing items or scheduling deliveries). Each agent in the toolchain is likely produced by a different organization since each serves radically different functions.

In multi-agent systems like these, agents must behave appropriately when interacting with humans and when interacting with other AI agents. If they can engage in conflict, like in Sugarscape, disastrous outcomes abound. A road full of aggressive self-driving cars presents a danger to passengers and bystanders, and an automated shopping system may order objectionable inventory if the individual agents in the system do not engage in ethical analysis of the user's commands.

Further, agents in a system may have contradictory aims. When confronted with this situation, the agents involved must act for the betterment of humanity. The most significant conflict environment in which AI agents must engage in ethical reasoning is warfare.

AI agents are deployed into military conflicts whether through automated drones, conflict zone analysis tools, or as defensive and offensive agents in cybersecurity. These AI agents come from a variety of manufacturers and developers, yet they must coexist with human operators and fellow AI agents to identify enemy combatants (human or artificial). They must engage in explicit ethical reasoning to analyze their situation through the lens of just war theory, internationally recognized law, and national military codes.

In combat, time is of the essence to make quick tactical decisions. However, choosing hastily or selfishly can lead to disastrous outcomes with horrific collateral damage to non-combatants. If these AI agents adhered to well-understood ethical principles in which war is morally permissible (such as Aristotelian virtue ethics), then legal scholars and philosophers can reason about how and why these agents make choices in the heat of combat.

VIII. Reproducibility of Experimentation

Our Sugarscape implementation is deterministic which means our results are eminently reproducible.[1] Software requirements are found in the `README` within the software repository. We used the `v2025.1` release. Other versions will produce similar, but not identical, results. We provide the JavaScript Object Notation (JSON) configuration files from our experiments.[2] The configurations are stored in the `conflict-ubi.zip` archive. The simulation takes a configuration file as input. To run a single simulation of Sugarscape with a configuration, run the following (substituting the system's Python 3 alias):

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

To effectively rerun the presented dataset, the top-level repository configuration file `config.json` must be modified. The `numParallelSimJobs` property specifies the number of simulation instances allowed to run in parallel. This should be scaled to the number of CPU cores or hardware threads available leaving enough to effectively run the operating system. Our Sugarscape implementation is memory-efficient, so the data collection bottleneck is available cores. The unzipped configuration files should be placed in the repository's `data` directory before execution.

Run `make setup` before data collection to check for a Python 3 installation. This will update values in the `Makefile` and `config.json` accordingly. Run `make data` to perform data collection. Run `make plots` to generate graphs (`matplotlib` must be installed). Tinkering with the script in the `plots` directory is necessary to produce exact formatting of the graphs presented in this work. To generate the dataset and produce graphs, run:

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

Since local data collection will likely be slow, the process can be readily extended for distributed execution. For instance, we used the Open Science Grid [29] and the HTCondor batch system [30] for data collection. A management script can be written to handle job creation and submission, regardless of the batch system used. We encourage this approach to speed up results reproduction.

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

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

IX. Conclusions and Future Work

We have demonstrated the necessity for AI agents to behave cooperatively when faced with conflict. In Sugarscape, we dialed up the amount of conflict possible in the simulation as a means of making the environment more treacherous. From previous work [5], we saw that social policies like universal basic income could serve as a useful surrogate for ethical behavior. However, the benefits of UBI were eroded when marginal increases in agent conflict arose.

Sugarscape societies were far more likely to succeed with the cooperative, utilitarian approach. Agents lived longer and engaged in less combat in societies which were far more likely to survive. Our results demonstrate the necessity for AI agents to engage in cooperation via ethical reasoning over other approximate substitutions like UBI.

There are a number of fruitful avenues for future work. Most apparent is the translation of other ethical theories into agent decisionmaking algorithms, such as Aristotelian virtue ethics. We do not claim that utilitarianism presents some definitive best set of ethical principles, so it is important to explore alternative approaches. There is a significant amount of work to be done when making a faithful translation of the originating philosopher's intent, but once completed one could compare the performance of a large stable of *algorithmified* ethical theories in the face of high-conflict, multi-agent AI systems.

A separate avenue of future work continues the investigation presented here. Since we have demonstrated the importance of cooperative, ethical behavior, we could determine at what point a cooperative AI system breaks when it is infiltrated by greedy, destructive agents. This has relevance across a number of AI applications such as the arms race escalation between generative AI tools and defensive AI tools used to poison and trap generative models when scraping works of art or copyrighted publications online.

X. Acknowledgments

We thank Ankur Gupta for his guidance in the refinement of our UBI mechanism and his review of our experimental methodology. This work was facilitated in part by using services provided by the Open Science Grid Consortium, which is supported by National Science Foundation awards #2030508 and #1836650.

References

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

[2] 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.

[3] 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.

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

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

[6] A. Yang, *The war on normal people: The truth about America's disappearing jobs and why universal basic income is our future*. Hachette UK, 2018.

[7] D. Graeber, *Bullshit Jobs: A Theory*. Simon & Schuster, 2018.

[8] R. Hasdell, "What we know about universal basic income: a cross-synthesis of reviews," *Stanford Basic Income Lab*, 2020.

[9] A. J. Auerbach and L. J. Kotlikoff, "Evaluating fiscal policy with a dynamic simulation model," *The American Economic Review*, vol. 77, no. 2, pp. 49–55, 1987.

[10] R. R. McDaniel, R. S. Sullivan, and J. R. Wilson, "A simulation model for welfare policy analysis," *Socio-Economic Planning Sciences*, vol. 22, no. 4, pp. 157–165, 1988.

[11] A. V. D. Luduvice, "The macroeconomic effects of universal basic income programs," *Journal of Monetary Economics*, p. 103615, 2024.

[12] F. Squazzoni, W. Jager, and B. Edmonds, "Social simulation in the social sciences: A brief overview," *Social Science Computer Review*, vol. 32, no. 3, pp. 279–294, 2014.

[13] G. Gilbert and J. Doran, *Simulating Societies: the computer simulation of social phenomena*. UCL Press, 1994.

[14] J. Mitcham, "Agent-based simulation of police funding tradeoffs through the lens of legitimacy and hardship," *Journal of Artificial Societies and Social Simulation*, vol. 26, no. 3, 2023.

[15] F. Neves, P. Campos, and S. Silva, "Innovation and employment: an agent-based approach," *Journal of Artificial Societies and Social Simulation*, vol. 22, no. 1, 2019.

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

[17] 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.

[18] 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.

[19] 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.

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

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

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

[23] 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.

[24] W. Wallach and C. Allen, *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[25] D. Leben, *Ethics for robots: How to design a moral algorithm*. Routledge, 2018.

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

[27] T. Hobbes, *Leviathan or The Matter, Forme and Power of a Commonwealth Ecclesiasticall and Civil*. Printed for A. Crooke, 1651.

[28] J. Sachs *et al.*, *Nicomachean Ethics*. Hackett Publishing, 2011.

[29] 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.

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