IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 October 2024, accepted 12 December 2024, date of publication 17 December 2024,
date of current version 30 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3519139

RESEARCH ARTICLE

Blueprints for Machine Ethics: A Digital Terrarium for Socio-Ethical Artificial Agent Decisionmaking

NATHANIEL KREMER-HERMAN[1], (Member, IEEE), ANKUR GUPTA[2], (Member, IEEE), AND ERIC R. SEVERSON[3]

[1]Department of Computer Science, Seattle University, Seattle, WA 98122, USA
[2]Department of Computer Science and Software Engineering, Butler University, Indianapolis, IN 46208, USA
[3]Department of Philosophy, Seattle University, Seattle, WA 98122, USA

Corresponding author: Nathaniel Kremer-Herman (nkh@seattleu.edu)

This work was supported in part by the Seattle University Undergraduate Summer Research Award and by the Butler University Holcolm Award Research Fund. 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.

ABSTRACT In a world increasingly driven by artificially intelligent autonomous agents, it is imperative that these agents behave according to well-understood ethical standards. We provide a unified computational framework for simulating societies of autonomous agents and for evaluating the impact of individual agent decisions on overall societal success. To the best of our knowledge, this *digital terrarium* is the first such system that allows for a direct comparison of various socio-ethical agent behaviors. As a first step, we present implementations of three popular ethical theories: Jeremy Bentham's hedonic act utilitarianism, ethical altruism, and ethical egoism. We compare these algorithmic *decision models* in a head-to-head manner and demonstrate that cooperative, utilitarian societies lead to vastly superior societal outcomes compared to the other two. Our explicit goal is to demonstrate the overt benefit of transparent, algorithmic ethical decisionmaking, in which developers, vendors, and users operate with awareness of the forms of ethical reasoning utilized by their technologies. Our results are a validation of a rich, interdisciplinary tapestry of work concerning the benefits of cooperation versus greediness. We enhance our results by also considering the impact of individual agent *greed* on a society's overall success (or failure). Though exploring greed (and similar phenomena) is interesting in itself, our more important contribution is our malleable, robust, and foundational experimental framework that facilitates comparative insights across a wide variety of popular socio-ethical models. We invite scholars from all disciplines to use our digital terrarium to grapple with challenging questions using this new simulation-focused methodology, in hopes that it may meaningfully shed some light on these questions. Our experiments in this article are not exhaustive. Rather, they are illustrative of the role that simulation can play in answering or understanding big questions in our society in response to rapid technological change.

INDEX TERMS Machine ethics, artificial intelligence, agent-based modeling, computational cultural modeling, cooperative systems.

I. INTRODUCTION

Humans are increasingly affected by autonomously-acting, artificially intelligent (AI) computer systems making decisions on their behalf. Many of these *algorithmic* decisions are seemingly inconsequential or benign (such as AI customer service systems), but other applications have more significant impacts (such as self-driving cars). These automated systems are placed in scenarios where they are tasked with making decisions with ethical ramifications. These decisions are frequently analyzed and enacted without any immediate awareness by programmers, vendors, or users. The outcomes may only be observed by people long after the inciting action was made. What makes these scenarios challenging is codifying how one makes an *optimal* decision, especially when navigating potentially contradictory preferences.

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad J. Abdel-Rahman.

We expect artificially intelligent agents to make decisions according to some set of values (and associated priorities) that are defined as optimal. Effective ethical decisionmaking involves prioritizing the right choice between competing actions in response to some dilemma [1]. The prioritization of actions is done according to the values imbued in the agent, which may come from a specific school of thought or a blend of many. Our work emphasizes measurable outcomes of societal wellbeing from the consequentialist school of thought.

Different normative ethical theories (such as utilitarianism, Kantian deontology, Stoic ethics, etc.) provide frameworks by which the set of possible agent actions are evaluated. Algorithms can be developed to approximate this ordering of actions, just as ethicists attempt to order human priorities in the making of decisions. This in turn produces an internally consistent way to rank or provide a partial ordering of those actions. For example, we could order according to moral duty, the cultivation of virtue, adherence to contracts and rules, or some other metric.

What is absent as artificial intelligence increases in power, breadth, and adoption is a true grounding of an autonomous computer system's decisionmaking process in well-known ethical theories. In many cases, such attempts are purely theoretical, shallow, or are overshadowed by competing constraints (e.g. profit motives, ease of implementation, or practical concerns), rather than exploring the nuances and implications of a robust decisionmaking process. If intricate autonomous systems for self-driving cars and other applications are indeed basing their decisionmaking on ethical theory, their designers have been reticent to submit this work for peer review. Neither have decisionmaking algorithms, ethical or otherwise, been regularly shared as free and open source software; doing so would allow for healthy public debate about the logic of machine decisions. The current body of publicly available work turns up scant few results about this important aspect of computing. Our work provides a humble step toward bridging this gap.

Even considering the external influences on the design of an ethical autonomous agent, there are concrete advantages to embedding an unadulterated implementation of an ethical theory in a computational context. Since such a system would be transparent, it would not suffer from the typical criticisms [2], [3], [4], [5] that are levied on a *black box* technology. That same transparency would allow an easy pipeline for human verification (either by reasoning from first principles or reverse engineering the output back to the source). Additionally, humans can apply their intuition to see whether the decisions made by a transparent autonomous agent are sensible. For example, one can reason about whether an agent's decision adheres to Kantian deontology, since that model of behavior is well-understood by a typical scholar of ethics.

We engage in this work with a crucial disclaimer: our implementations of socio-ethical theories are offered here at a layperson's level of understanding both for straightforward implementation and for ease of comprehension for a diverse audience. As such, we invite critique and proposed enhancements to our understanding and implementation of these theories. We make the bold (and, for our approach, necessary) assumption that many aspects of ethical, socially-considerate behavior are computable, either exactly or by a high fidelity approximation. While we do not rigorously test that assumption here, we leave the door open for future cross-disciplinary research to inform this strategy.

It is incumbent upon us to leverage the work of philosophers and see how those ideas apply to our modern, computational landscape. We construct a universal computational platform upon which we can compare different ethical (and social) decisionmaking models. We provide useful context to practitioners who seek advice on identifying the most appropriate agent decision model for hands-on applications. We do not assume that an ethical decision model adopted for a particular application will be universally appropriate for all artificially intelligent products. Rather, we seek transparency about the decisionmaking process of machines and believe ethical theories can, at minimum, provide guidelines for artificial decisionmakers performing a specific task.

For the sake of simplicity, we refer to our computational platform as a *digital terrarium*, where we introduce agents that interact with a simplified model of the real world. Our digital terrarium is inspired by Epstein and Axtell's *Sugarscape* [6], an agent-based simulation on a grid world. In previous work, we provide a complete implementation of Sugarscape [7]. Our implementation also allows for independent verification of results according to modern software development standards.

II. OUTLINE OF ARTICLE

Our work has many inter-related components, reinforced both by quantitative experiments as well as qualitative reflection. In this section, we provide a high-level road map to assist the reader in navigating the various facets of our work. In Section IV, we describe *Sugarscape* [6], [7] and explain its baseline feature set. We then describe an agent-level decision layer in Section V that supports more nuanced decisionmaking than the default behavior of Sugarscape. In particular, we implement the *utility calculus* [8] as our first direct use of this decision layer structure.

Section VII describes the bulk of our experimental results across two styles of investigation: (1) the relative societal success of different decision models (Section VII-A) and (2) the importance of prioritizing self versus society (Sections VII-B–VII-D). We also spend some time discussing the critical software engineering principles of fidelity and reproducibility of our results in Section IX. We additionally provide a more granular outline with contextualized discussion.

A. SUMMARY OF AGENT DECISION LAYER

We add a *decision layer* atop standard Sugarscape in Section V. This replaces an agent's default, greedy behavior

with a specified decision model. Our emphasis in this paper is to study ethical decision models, but one could conceivably implement models for other domains such as biology, economics, psychology, or sociology.

Our digital terrarium allows us to study trends within the artificial society and assess the successes and failures of agent behavior. Epstein and Axtell [6] observed interesting societal behavior even when individual agents acted in a greedy fashion. Our goal is to capture *emergent behavior*: the behavior of groups of agents (or a society) that transcend the available actions of any individual agent in the group.

The first ethical decision model we implement in our decision layer is *hedonic act utilitarianism*, as introduced by Jeremy Bentham [8]. The details of this algorithm are presented fully in Section VI, along with carefully-considered nuance about how to best preserve the spirit of the original utility calculus. An act utilitarian decisionmaker considers the consequences of any potential action across all affected agents and aggregates the overall good and bad consequences of that action for all stakeholders. Its ultimate goal is to choose the action that produces the greatest good for the greatest number of agents.

Utilitarianism is, in some sense, quite appealing: it promises an objective, non-egocentric, and in our context, computable, way to evaluate the ethical quality of a proposed action. This procedure roughly corresponds to the traditional computer science notion of rational agent behavior, where the goal is to maximize one's expected utility. The commonality between the ideas is striking, and implementing act utilitarianism is an ideal first step to illustrate the viability of our experimental framework.

We provide a software representation of Jeremy Bentham's pseudo-algorithmic *hedonic calculus* which seeks to be faithful to the source text. We further implement two additional extensions of the hedonic calculus: *ethical altruism* and *ethical egoism*. Practitioners of altruism engage in pure selflessness while egoists engage in pure selfishness whereas utilitarians act in an egalitarian manner.

B. SUMMARY OF THE DIGITAL TERRARIUM

Our primary technical contribution is our digital terrarium for comparing socio-ethical decision models. Figure 1 shows the software architecture of our contributions. The foundation of our digital terrarium is our implementation of the Sugarscape agent-based model. This simulation framework creates an environment where individual, autonomous agents (i.e. AI agents) interact at the micro-level which leads to macro-level emergent behaviors in the artificial society.

Atop Sugarscape we build a decision layer which allows drop-in replacement of the default (and greedy) agent behavior with more nuanced socio-ethical decision models. In particular, we showcase utilitarianism as our first decision model. We also implement ethical altruism and ethical egoism as distinct ethical models which are derived from utilitarianism. These models and their implementations are discussed in Section VI-C. When executed with these decision models enabled, Sugarscape provides a society of autonomous agents which each adhere to one of these ethical decision models.

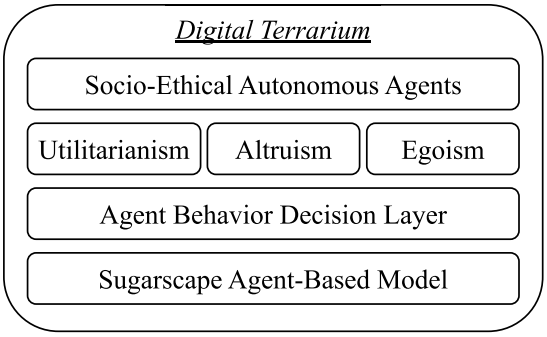


FIGURE 1. Digital terrarium software architecture.

With this software architecture in place, we perform a variety of experiments to validate our implementation of these socio-ethical decision models and demonstrate their relative effectiveness. We first compare utilitarianism (as a cooperative approach to decisionmaking) to altruism and egoism which contain inherent flaws related to the amount of greed present in altruist and egoist agents. These flaws are described in Section VI-C. We then determine how greedy society is allowed to be and how greedy individual agents are allowed to be and still yield stable, prosperous societies.

C. SUMMARY OF AGENT GREED

In our digital terrarium, we demonstrate that societies where agents act in a utilitarian manner definitively outperform societies where agents act in their own best (greedy) interests and societies where agents engage in wasteful self-sacrifice. In particular, utilitarian societies: are *more than twice as likely* to survive any given initial landscape, gather *thrice as many* resources as the other societies, and are composed of agents that live *longer and happier* lives than agents in other societies. These findings imply that utilitarian societies (and the agents in those societies) are healthier, more productive, and more fully utilize their environment than their counterparts adhering to the other implemented ethical theories.

We reinforce the computational use of utilitarianism as a guiding ethical decision model through further consideration of two implementations of greed. In the first (described in Section VII-C), we add an additional property to an agent's decisionmaking: their *selfishness factor*. Societies composed of agents that are more than 80% selfish lead to volatile societal outcomes (and often lead to extinction). Purely selfless societies of agents who are 0% selfish are likewise volatile and prone to failure. Utilitarian societies (at or near 50% selfish) are incredibly successful, and the probability of success is nearly guaranteed.

In Section VII-D, we describe the second consideration of greed, where we consider societies of agents with a mixture of decision models (namely the ethical egoist and

utilitarian models) and compare outcomes for societies which are purely egoist, purely utilitarian, and somewhere in-between. We note an inflection point showing a society of at least 40% utilitarian agents begins to produce the successful societal outcomes of our other experiments. Further increasing the percentage of utilitarian agents in the society reduces the volatility of outcomes and reinforces the success experienced. The sum total of our results are extremely promising and are a compelling first step toward a fully realized, formal exploration of computational ethical decision modeling.

D. SUMMARY OF RESULTS

Our work introduces multiple kinds of outcomes: software, experimental results, and interdisciplinary discussion. We distill our results here for clarity. We provide the following key deliverables:

- Digital terrarium for evaluating socio-ethical decision models for autonomous agents (Section VII)
- Algorithmic representation of Jeremy Bentham's hedonic act utilitarianism (Section VI)
- Experimental validation that cooperative societies vastly outperform both purely selfish and purely selfless societies across a variety of metrics. (Section VII-A)
- Experimental validation that some limited level of selfishness is tolerable in an otherwise cooperative society across two different ways of representing selfishness (Section VII-C)
- Certification of reproducibility for all claimed results (Section IX)
- Identification of rich, interdisciplinary opportunities for collaborative future work (Section X)
- Discussion on the computational tractability of ethics and ethical autonomous agents (Section X-B)

These deliverables demonstrate a comprehensive evaluation of our digital terrarium. We also show how one can exactly reproduce our findings. Our novel results are individual threads which are woven into an already vast tapestry of rich, interdisciplinary work which we present with care.

We argue that algorithmic decisionmaking should be developed transparently instead of relying upon the preferences of private firms to determine how machines make socio-ethical decisions. We start with Bentham because his approach is simple and straightforward, not because he is in any sense correct or definitive. Upon injecting Bentham's approach into Sugarscape, we find that good things happen to society. Therefore, we argue developers of artificially intelligent agents should be obligated to share their decisionmaking algorithms and what moral understandings govern their design choices. One can do much worse than the early-modern simple utilitarian calculus provided in our treatment.

To the best of our knowledge, we provide the first *algorithmified* realization of Jeremy Bentham's hedonic act utilitarianism for autonomously-acting computer agents. We draw directly from Bentham's work and apply it to our digital terrarium. Our work is exciting because it has wide-reaching implications across many domains as we attempt to answer big questions in the humanities by way of a very different methodology. In particular, our work is both novel and intellectually satisfying because there do not appear to be any other solutions, methods, or standards by which we can benchmark our implementation other than by qualitative comparison or by analogy.

III. RELATED WORK

We provide the typical, broad scholarly overview of related work relevant to our digital terrarium. Since our work is both novel and highly interdisciplinary, we also provide an abbreviated survey of works relevant in other domains which may be compared to our work in a qualitative manner. Our work is inherently interdisciplinary, and it would be impossible to express a full accounting of all relevant research. We provide this survey for the astute reader of these domains, acknowledging that we make many simplifications in this work which allow us to focus, with clarity, on the computer science perspective we present. By referring to the bodies of work, we demonstrate the simplifications made here are intentional, informed, and aim to convince the reader of the sincerity of our study and of the validity of our results.

A. ABBREVIATED SURVEY OF INSPIRATIONAL WORK

In the spirit of *Appendix N* [9], we provide a high-level, incomplete list of related works which have inspired us. We offer these as suggestions for the astute reader and humbly acknowledge that the referenced works mark only the *beginning* of an interdisciplinary journey. As a top-level observation, we note that many lines of inquiry among disciplines relevant to our digital terrarium pose intriguing theoretical observations. We submit that our digital terrarium is a unified playground for any overlapping field of study to experiment, and our contributions are purely *additive* to the existing body of work.

We make the assumption that artificial intelligence affects and is affected by practically all other disciplines. This assumption is reinforced by the mass adoption and rapid innovation of artificial intelligence across disciplines and professions not unlike the adoption of the World Wide Web and of the personal computer. We divide fields of study broadly based on their placement in the traditional liberal arts and mechanical arts which have a storied history of influence upon each other.

Our digital terrarium ingests the theoretical *observations* relevant to ethical, artificially intelligent agents gathered across the liberal arts and the *applications* for ethical AI agents from the various mechanical arts. We provide *blueprints* for the creation and deployment of ethical AI agents which can be fed into various AI solutions. These ethical agents are intended to spur on future observations and applications which create a virtuous cycle of progressively

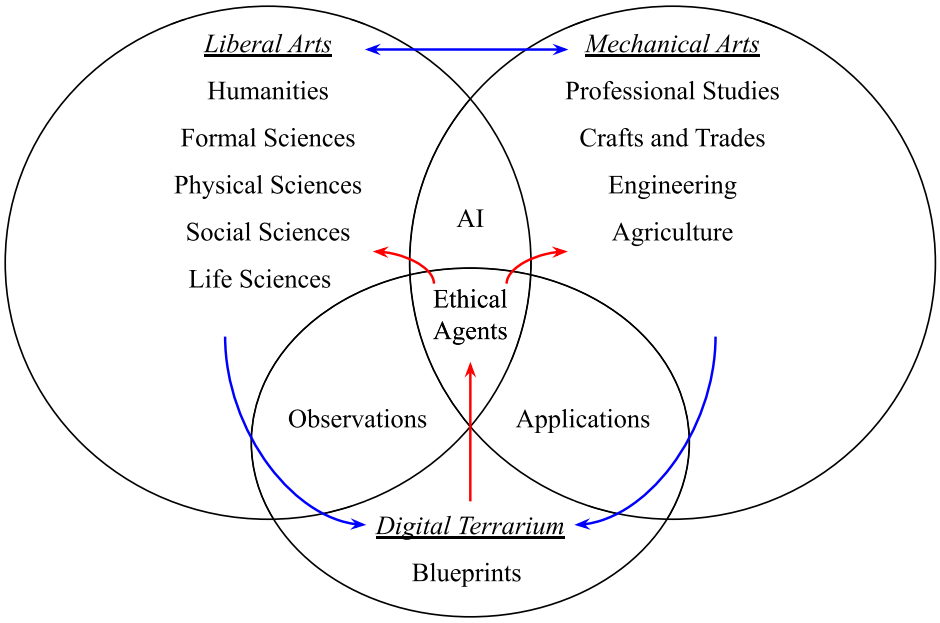


FIGURE 2. Interdisciplinary connectivity of our digital terrarium.

more optimal, and more ethical, AI agents. Figure 2 provides a diagrammatic representation of the connectivity of our digital terrarium to other disciplines at coarse granularity.

The ethical agents under study in our digital terrarium are the result of interdisciplinary considerations. We do not claim to overshadow any field of study but rather make a good faith effort to provide a useful application of their respective bodies of knowledge that synthesizes lessons learned across disciplines. We add a computational framework for these fields to continue positively affecting the development of ethical artificially intelligent agents in our global society.

Especially relevant, and likewise inherently interdisciplinary, fields of study include machine ethics, artificial life, game theory, and complex systems. We place these fields in the same area of the Venn diagram in Figure 2 as artificial intelligence with the same significant, reciprocal impacts upon the liberal and mechanical arts. Through our digital terrarium, we hope scholars in these fields can meaningfully explore ethical agent decisionmaking in complex artificial societies.

B. SURVEY OF COMPUTATIONAL MODELING AND AGENT BEHAVIOR

Modeling techniques for digital life simulations have a long lineage, going as far back as John von Neumann's seminal work on self-reproducing automata [10]. Langton [11] helped found the study of artificial life. One of the first fruitful attempts at applying agent-based simulation techniques for social phenomena was by Schelling [12]. He pursued a wide body of work along those lines but was fundamentally constrained by available computational power. Our work extends the agent-based simulation *Sugarscape* [6], an evolution of Schelling's work additionally incorporating ideas from Conway's Game of Life [13]. More recently, Jara-Ettinger et al. [14] implement a complex Markov decision process (MDP) that allows agents to both navigate toward their goals as well as have some choice about which goals to pursue. Their work is focused on identifying an unknown (but fixed) agent decision model rather than observing emergent behavior like Epstein and Axtell's Sugarscape [6].

When deploying ethical agents, we must consider how and in what way to incorporate ethics; the following works attempt to categorize either the end product or the process by which ethics is evaluated. Winfield et al. [15] outlines four categories for deployed ethical agents: *ethical impact* (agents that can be evaluated for ethical consequences), *implicitly ethical* (agents designed to avoid certain unethical outcomes), *explicitly ethical* (agents that reason about ethics), or *full ethical* (agents that make and can justify specific ethical judgements). Segun [16] argues that one should carefully consider ethics at all points during artificial agent creation: during ideation, development of pseudocode, and implementation of a real agent system. The ethical focus and concern is different at each stage of development, but each is equally important to the overall process.

Branches of normative ethics (such as consequentialism, deontology, and virtue ethics) offer opportunities for experimental implementation for autonomous agents in various ways [17]. An early report in computationally evaluating an ethical dilemma is presented by McLaren [18] comparing approaches from the Truth-Teller and SIROCCO systems. One can also model public policy in an artificial society as presented by Diallo et al. [19]. The GenEth dilemma analyzer [20] is a knowledge representation scheme for

identifying ethical features of actions (particularly the presence or absence of those features) in a human-in-the-loop model. Each of these approaches have differing use cases and implementations, yet they all provide meaningful examples of implementing ethics computationally.

Many ethical theories emphasize the importance of the good of all as opposed to the good of solely the self. This matter of greed versus generosity has been studied across many disciplines including philosophy [21], psychology [22], [23], [24], [25], [26], [27], business [28], [29], and biology [30], [31], [32], [33]. Williams [34] provides argument that some degree of greed is essential for human welfare. Cassill and Watkins [35] reinforce this by clarifying the importance of greed when agents are presented with resource scarcity. Indeed, one would be hard-pressed to find a field of study which does *not* have something to say about greediness and generosity. These works investigate competing characterizations of greed and often distinguish between *self-interest* and *greed*. We do not make any such distinction but instead focus on how agents' self-importance (or lack thereof) in their ethical decisionmaking leads to emergent societal outcomes.

Agent-based modeling allows investigating not just digital life but also ethical modeling [36]. However, previous attempts to do so have neglected to take into account the full breadth of ethical issues and to fully evaluate their results [37]. Our digital terrarium is an agent-based simulation model for computational social sciences that overcomes these criticisms. Most directly related to our work, Lasquety-Reyes [38] extends Sugarscape to explore algorithmic virtue ethics. He implements *temperance* from the Aristotelian virtue ethics tradition with a complex quantitative formulation using the PECS (physical, emotional, cognitive, and social) model to moderate these differing agent desires. Lasquety-Reyes provides a tentative first step toward a fully realized implementation of virtue ethics in code.

C. SURVEY OF THE ETHICS OF ARTIFICIAL INTELLIGENCE AND MACHINE ETHICS

Ethical issues in artificial intelligence range from concerns regarding the adoption of AI technologies, ethical use cases for AI, governance and policy surrounding AI, and how to create ethically-behaving artificial agents (machine ethics). Our work expands the field of machine ethics. We first provide introductory material on the ethics of artificial intelligence which serve as relevant background information.

Smith and Browne [39] provide a walkthrough of many socio-ethical issues involving computer technology and society with a primary audience of business management and the general public. They dedicate significant time to artificial intelligence and AI decisionmakers, and they discuss their impact on the workforce and society. Coeckelbergh [40] presents an accessible summary of many ethical concerns with the adoption of AI and does so in a way that nicely sidesteps the more overly dramatized aspects of the topic. Boddington [41] provides a formal introduction to AI ethics. Floridi [42] discusses, among other illuminating ideas, that the importance of ethical design must increase in no small part due to the growing distance between agency and intelligence in automated agents. Eubanks [43] discusses algorithmic bias and the impacts of automated systems on the socioeconomically disadvantaged, compounding their struggles. Her focus on algorithmic injustice has strong ties to topics covered in other relevant works. Liao's compiled volume [44] grapples with challenges of determining the preferences needed for artificial agents to mimic human behavior, the transparency (or black box nature) of AI, planning for a universal basic income, and the ethics of an AI lover, among others.

Quinn [45] provides an overview of many popular ethical theories, their application to (computer) technology issues, and a plethora of case studies covering the entanglement of modern societal issues with rapid technological evolution. His computer science background grounds the issues in a way that is appropriate for general readers and is especially insightful for computing professionals. Kearns and Roth [46] provide a socio-ethical overview of concerns arising from machine learning and from an increasingly algorithmified world. The authors leverage their computer science expertise to provide general audiences an understanding of what goes on *under the hood* with complex algorithmic processes, such as mass data collection and resulting (privacy breaking) advertising.

The body of work surrounding machine ethics and similar terms is *vast*. We provide some highlights which range from general audience appeal to specific philosophical works. These highlights serve as introductory matter to further reading and as demonstration of our appreciation and care for the many thorny issues surrounding machine ethics, many of which we simplify for clarity of our contributions.

Blackman [47] provides a treatment on artificial intelligence and machine learning, which is catered toward business leaders although it is also an appropriate introduction for general audiences to many of the socio-ethical concerns surrounding AI. Leben [48] provides an engineer-focused treatment of concerns and strategies for writing ethics algorithms for machines. He focuses on the fact that machines do not carry the moral psychology embedded in modern humans and makes the case that a Rawlsian contractarian approach may be the most straightforward for ethical machines. Wallach and Allen [49] similarly provide an overview of machine ethics, covering the primary concerns one should consider when designing ethical machines. Like Leben, they dive into considerable philosophical content to prepare engineers for the challenges of working in the machine ethics space. Extending that theme, Lin et al. [50] compile a number of interesting articles that extend that overview to ethical considerations for both the software of AI as well as its robotic counterpart.

The previous works have been, to varying degrees, covering machine ethics for the non-expert. Anderson and Anderson [51] put together one particularly thorough compilation primarily written by philosophers and for other philosophers as their audience. They not only provide us with grounding for the term *machine ethics* but cover a broad range of relevant issues in the field including the moral agency of artificial decisionmakers, the distinction between machines made to follow ethical principles and those which derive ethical principles, human-robot interaction, etc. This work can claim a significant share of the reason that machine ethics has become established as its own line of inquiry.

D. SURVEY OF GAME THEORY

For both the Sugarscape simulation and Jeremy Bentham's utilitarianism, there is a clear link to economics and game theory. The concept of utility functions shares an etymological origin with Bentham's conception of hedonic utility, and Bentham's emphasis on rational, detached decisionmaking also makes clear the connections between game theory and utilitarianism. At its core, utilitarianism strikes a balance between extremely greedy behavior (pure self-interest) and extremely sacrificial behavior (pure selflessness). One of the most popular game theoretic approaches to exploring self-interest is the Prisoner's Dilemma [52].

Axelrod's investigations into the Prisoner's Dilemma [53], [54], [55] are perhaps the most well-known of these. His inherently interdisciplinary work connects political science, behavioral science, economics, mathematics, and philosophy. The utility calculations of the players in a Prisoner's Dilemma are similar in nature but quite different in implementation to Bentham's utilitarianism. We explore the relationship between self-interest and societal success in a similarly interdisciplinary fashion but do not constrain ourselves to the limited game rules of the Prisoner's Dilemma. With Sugarscape, the set of possible agent actions is much larger, the environment is more interactive, and there are far more agents involved.

We also note the relevance of evolutionary game theory to our work. The field has a rich and influential history [56], [57], [58]. One key approach to evolutionary game theory is evolutionarily stable strategy analysis [58]. This is essentially a Nash equilibrium [59] with an added stability criterion given each agent has a fixed strategy. It provides a deterministic, *if-else* action selection strategy for an agent's interactions with others. A second approach to evolutionary game theory is evolutionary dynamics [60]. This approach considers the possibility that an agent can change their decision strategy to maximize its fitness. This is conceptually similar to Jeremy Bentham's goal of people maximizing their utility [8] which we attempt to replicate in this work.

E. SURVEY OF COMPLEX SYSTEMS, ARTIFICIAL LIFE, AND THE SUGARSCAPE MODEL

The Sugarscape simulation we use is an example of a *complex system*. From large networks of autonomous moral agents to a simulated brain cogitating on moral dilemmas, complex systems research has significant ties to machine ethics. Ladyman and Wiesner [61] provide an extensive treatment of complexity and provide many examples of complex systems in order to define the term itself. While many of these examples model the physical sciences, some are related to artificial life and computational social science (including Sugarscape). The study of artificial life has particular relevance to our work, and we provide highlights of related works within this subfield of complex systems to demonstrate our consideration of complexity as it pertains to machine ethics.

Artificial life (which includes artificial societies like Sugarscape) is an inherently interdisciplinary field and has been since its inception [11]. We provide some relevant works in artificial life to our work. Complexity appears at the level of neurons in an artificial agent's brain, the collection of agents' expressed behaviors from those neurons interacting, and finally the emergent societal behavior [62]. An agent's behavior can also be influenced by querying the mental state of other agents through communication [63]. This can create a feedback loop of agents affecting each other. At the societal level, agents may make commitments to each other [64] (with implicit or explicit worth given to those agreements). They may also share death stories with their peers [65] with the goal of increasing the odds of societal survival.

The study of artificial life also has ties to machine ethics with open questions remaining on the agency and moral rights of artificial agents [66]. We explore ethical issues in this work, and we do so through artificial life (in our case, an artificial society). However, we focus on observations made from observing ethical decisionmaking made by agents and eschew questions related to agency and moral rights.

Čejková [67] presents a compilation of works on artificial life and a modernized English translation of Karel Čapek's play *R.U.R. (Rossum's Universal Robots)* in celebration of the play's centenary. *R.U.R.* has the distinction of introducing the word *robot* to the English language and provides many insightful critiques of the social consequences of unfettered automation which are incredibly relevant to the mass adoption of artificially intelligent technologies. Additionally, Čapek explores questions surrounding the agency and inherent worth of the robots in his play (which, or who, are essentially treated as slaves). Čejková's addition of essays reflecting on *R.U.R.*'s impact upon various branches of artificial life demonstrates the inherently interdisciplinary nature of the field.

Sugarscape has its origin in computational social science and provides complex artificial societies one can use to reproduce real societal phenomena from individual agent behaviors. The source material [6] as well as follow-on work [68] ground the simulation in fundamental principles to the social sciences. As such, Sugarscape has effectively been applied to a number of social problems. It has been used to demonstrate the effectiveness of pensions and social security programs [69]. The simulation has also shown

TABLE 1. Popular sugarscape implementations and feature set provided by book chapters.

Implementation	Chapter 2	Chapter 3	Chapter 4	Chapter 5
NetLogo [74] Sugarscape	✓	X	X	X
MASON-Sugarscape [75]	✓	✓	✓	X
Our Implementation [7]	✓	✓	✓	✓

the effects of technology proliferation and resulting wealth disparities affecting the growth of societies [70], to examine tax structures [71], and investigate wealth adjustment and wealth disparities [72]. Another example involved adding a cognitive layer to agents [73], applying Sugarscape to psychology.

IV. OVERVIEW OF THE SUGARSCAPE MODEL

Sugarscape was originally developed in 1996 and evaluated in *Growing Artificial Societies* [6]. It is an agent-based simulation platform for exploring how societies might express emergent behavior based solely on the actions of individual agents. Sugarscape is a two-dimensional $n \times m$ toroidal grid landscape, where each grid cell contains some amount of the two available resources: *sugar* and *spice*. These resources are metaphors for wealth and are gathered by agents who then consume sugar and spice over time. These resources regenerate at each grid cell at (potentially different) predetermined rates. One can simulate barren lands or lands of plenty by adjusting a cell's regeneration rate.

Agents are born with a single, greedy goal: live as long as possible. Each timestep, an agent consumes sugar and spice according to their *metabolism* for each, and they can move as far as their *vision* allows to gather more resources. The simulation also contains additional features that can be toggled for both the environment (such as seasons, pollution, or disease) and agents (such as trade, combat, or reproduction).

Sugarscape provides an elegant baseline to study agent-based behavior, complete with rich social dynamics that range from simple, short-term behavioral rules to more elaborate long-term agent planning. It does, as Epstein and Axtell describe [6], a remarkably good job of reinforcing our understanding of real world phenomena and social behavior. The abstraction level both allows a metaphorical similarity with the real world, while not sacrificing the computability of discrete numerical values.

Figure 3 displays a single instance of our implementation of the Sugarscape model executed from start to finish. Most features are enabled including trade, combat, reproduction, disease, cultural transmission, and lending. Agents are colored red, sugar is yellow, spice is brown, and tan cells have both sugar and spice. The more saturated the color of a cell, the more resources are present at that location.

The top left panel shows the initial state of the simulation: a starting population of 250 agents are randomly placed across the environment. Some agents are placed in ideal, resource-rich locations while others begin life in more destitute areas. Among the agents, 50 of them are initially infected with a unique disease which can be spread to others through contact.

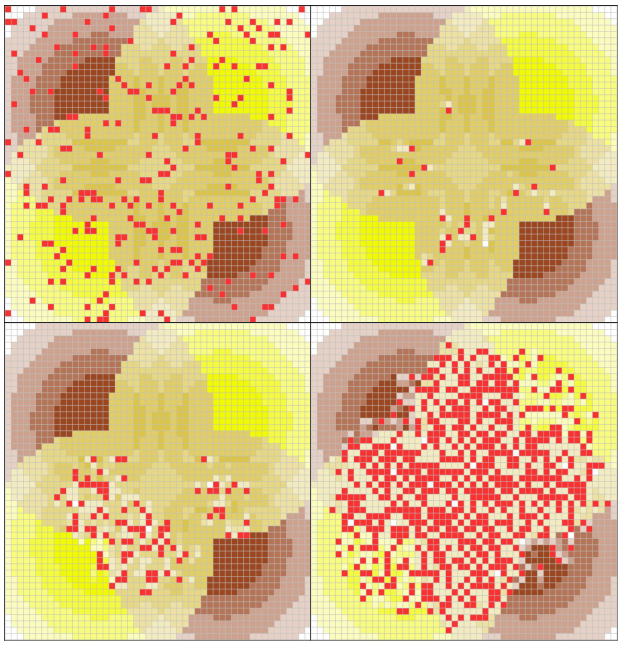


FIGURE 3. A sugarscape society from start to finish.

The top right panel shows the state of society 75 timesteps into the simulation run. Most agents have died due to selection, competition, combat, and disease. Societal extinction seems the most likely outcome, so it would appear the outlook for this society is bleak.

The bottom left panel depicts the society at 125 timesteps. We observe the beginnings of a population rebound. In this particular simulation run, disease was eradicated after it nearly wiped out the population. Note that the strongest growth is occurring in a particularly resource-rich zone with both sugar and spice readily available (which are both necessary for reproduction).

Finally, by timestep 200 shown in the bottom right, society has stabilized. They have survived the initial nosedive from harsh initial conditions and have overcome environmental challenges. This represents a successful society. Note the relative desaturation of the cells within and on the periphery of the population. This demonstrates the society is effectively using their available natural resources and is approaching their ecological carrying capacity.

Our ground-up reimplementation of Sugarscape is, to the best of our knowledge, the only open source, *complete* implementation of Sugarscape as presented in the source text [6]. In private conversation with its authors, there appears to be no such body of software available currently, though there have been partial attempts [74], [75]. Our version of

Sugarscape also includes robust validation, so any user of our software may verify the results presented here and in future work.

Table 1 compares two widely-used Sugarscape implementations with our version. We compare implemented features based on the chapters of *Growing Artificial Societies* [6]. Our implementation provides the full feature set from the book and additionally incorporates new features and mechanisms which make the validation of the book's results straightforward, provides greater extensibility of the mechanisms for future work, and incorporates simulation engine improvements in accordance with contemporary software development best practices. Previous work [7] provides a more complete description of our implementation's usefulness, its feature set, the reproduced results from the source text, and additional features which extend our version beyond the source text.

Sugarscape is an ideal platform for our work. Unlike other agent-based models, it is a fairly general-purpose simulation tool. The source text provides an interdisciplinary look into computational social science, and our extension of Sugarscape greatly broadens its applicability and multidisciplinary nature. In short: Sugarscape is a phenomenal agent-based model for investigating societal-scale research questions including how one can implement socio-ethical decisionmaking for autonomous agents.

V. AGENT DECISION LAYER

At the heart of our motivation in this work is the deep desire to better understand how to make ethical decisions computationally. This is an area of both technical interest in the data science and artificial intelligence communities as well as general broad interest due to its potential to affect human interactions. It is also of interest to scholars who may wish to study the comparative impact of different decisionmaking paradigms (be they social, ethical, or otherwise).

We extend our version of Sugarscape with a *decision layer* which agents use to govern their action prioritization. Our decision layer evaluates and scores all possible actions of an agent through a scoring function and decides which action the agent takes. That action is then executed by the agent in our digital terrarium. This approach exposes the nuances of ethical computational decisionmaking for careful study.

The default behavior of an agent in Sugarscape is greedy, where an agent will take the action that provides the most immediate total resources (the sum of sugar and spice). As one might suspect, it may not always be optimal in the long-term for an agent to maximize its resources in each timestep. There may be scenarios in which taking a seemingly suboptimal action now may yield an overall better resource gain.

This style of computational thinking intersects with artificial intelligence, where lookahead and accurate prediction may improve an individual's most successful action. However, in social and societal simulation systems such as Sugarscape, one might also want to incorporate the overall impact of other agents in the decision process as well. Our decision layer considers these differing models of computational decisionmaking with relative ease, including models that consider the preferences of others.

Agents can utilize the full power of visible[1] agent and environmental circumstances and assess their next move accordingly. In particular, this organization of our software provides a natural platform to make apples-to-apples comparisons among different decision models, from the very simple (greedy) to the more complex (cooperative). To study another decision model, one implements a new scoring function for the available actions according to the corresponding social or philosophical principles that govern that model.

Due to the rich feature set we implement in our digital terrarium, we provide a reasonably high fidelity representation of real human societies that is accessible, transparent, and inherently computable. One could add even more features to Sugarscape to further nuance the decision space: tracking and evaluating secondary or tertiary impacts of agent decisions, adding variables to study the impact of decision models on new social phenomena (such as population sprawl), or building an initial configuration to mimic a specific (perhaps real world) society. Each of these sample avenues of exploration is readily achievable within our framework.

Our digital terrarium serves as an initial means to benchmark decision models. Although such an evaluation is not truly a measure of real world success, considering decision models in this way may allow deeper thinking about the true impact of adopting a given socio-ethical theory in societal settings. We hope to inspire other researchers to think of our work as a hub for comparing and contrasting a diverse range of decisionmaking models.

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*

VI. BENTHAM'S HEDONIC CALCULUS

In our decision layer, we implement an algorithmified version of Jeremy Bentham's hedonic calculus (also known as the *felicific calculus* or the *utility calculus*), which forms the core of his ethical theory hedonic act utilitarianism [8]. To him, an ethical action is one that creates the greatest happiness for the most people. We provide a brief overview of Bentham's utilitarianism, but we will modernize the terminology for our computational context as necessary.

[1] We allow for, but do not presume, agent omniscience.

TABLE 2. A modern translation of Bentham's variables from the hedonic calculus.

Variable	Bentham's Definition	Sugarscape Terminology
Intensity	magnitude of happiness	raw sugar and spice gained
Duration	time that happiness is observed	extra survival time given intensity
Certainty	likelihood that happiness occurs	0 or 1, as actions are deterministic
Propinquity	time delay before happiness begins	0 or 1, as actions are immediate
Fecundity	likelihood happiness begets happiness	same as Bentham
Purity	likelihood happiness begets unhappiness	same as Bentham
Extent	number of agents affected by the action	same as Bentham

Bentham posits that an agent can calculate (either by observation or derivation) a quantitative value for a given possible action (representing its consequences). The decisionmaker then aggregates the scores for all agents affected by the potential action. He further declares that no agent is worth inherently more than any other.

Bentham refers to calculated positive values as *pleasures* and negative values as *pains*. We represent these ideas as a single variable: the *happiness* for a given action, which can take positive or negative values as appropriate. We define the *utility* (or inherent aggregate value) of an action as the amount of happiness gained or lost.[2] We assume (as Bentham does) that each agent prefers more happiness. The action chosen is the one that generates the most happiness across all affected agents and is thus deemed the most ethical decision.

Algorithm 1 shows pseudocode for the hedonic calculus, derived from the source text [8]. It has been updated slightly to conform to our modernized terminology. In this work, we deviate from the source text only to improve the speed of computation or in minor implementation details. We are careful to preserve the spirit of the original text. As such, it is a faithful representation of Bentham's hedonic calculus.

As one might imagine, calculating an agent's happiness is not particularly straightforward. Bentham also grappled with this problem; he approached it by identifying aspects of happiness (which he termed *circumstances*) that he felt are more easily quantified. While Bentham does not provide a formulaic way of combining these circumstances (as the computational language to describe this did not exist in his day), it is clear he believed happiness was quantifiable. We modernize his approach by thinking of these circumstances as *variables*, which, when quantified together, produce a formulaic representation of an agent's happiness. We use Bentham's variables to calculate happiness in line 6 of Algorithm 1.

We do not argue that Bentham's approach to quantifying happiness is in some way definitive. Bentham had no conceptualization of computation in the same way we do today, and in the centuries since his initial work was published scholars have iterated upon his ideas for quantifying happiness. As our first foray into socio-ethical decision models, we focus only on Bentham's thoughts as presented rather than attempt to capture the entire body of work in quantifying happiness.

A. THE VARIABLES OF BENTHAM'S HEDONIC CALCULUS

In our digital terrarium, we represent an agent's change in happiness as the relative gains of sugar and spice from taking a given action. Bentham defined seven variables to help quantify the utility associated with taking a given action: *intensity*, *duration*, *certainty*, *propinquity*,[3] *fecundity*, *purity*, and *extent*. In each timestep of our simulation, every agent computes our modernized interpretation of these variables prior to taking an action. We refactor Bentham's variables using appropriate modern computational terminology and describe them from an agent's point of view in Table 2.

Since an agent's happiness depends in part on the *extent* variable, each agent's utility is now entangled with the happiness of other agents in the digital terrarium. As such, a given agent's selected action must necessarily consider the overall happiness of nearby agents. Bentham defines fecundity and purity as having distinct characteristics from the other circumstances. He referred to these two circumstances as *properties* [8] which we interpret to mean that he struggled with quantifying these concepts. The closest computational interpretation of these two variables is as probabilities (i.e. the likelihood that the corresponding event occurs). However, in most cases, one must calculate the magnitude of future rewards to determine the desired probability. In truth, magnitude is the actionable information for these variables. As a result, our implementation predicts this future magnitude directly instead of via an indirect probability.

Implementing these variables and calculating an overall utility is not as easy as it might seem. Each variable interacts with the simulation state and the list of enabled features for a given simulation run in complex ways. Refactoring Bentham's variables for every combination of available features is one of the most challenging portions of this work. The intellectual depth of this aspect is significant and requires care and mathematical rigor to achieve sensible, consistent results.

Utilitarianism has great promise in computational contexts because of its inherent computability and its metaphorical similarity to much of the work in the artificial intelligence community. There are some complications with using utilitarianism carelessly, however. For example, this ethical theory implies that one can compare the utility of traditionally incomparable objects or ideas like freedom, art, a person's

[2] We define the utility lost as aggregate *unhappiness*.

[3] Also called *proximity* in the source text and in other referential works.

dignity, or the proof of a theorem. Utilitarianism also presumes that future consequences can be accurately defined (in terms of happiness), computed, and incorporated into a decisionmaking process *now*, which may not be a reasonable assumption for a particular simulation environment.

Practically speaking, resolving this assumption requires heuristics and becomes little more than an informed prediction as one approaches the horizon of computability. Importantly, scholars of ethics generally understand applying utilitarianism as a process of evolving estimations of ever-changing possibilities inherent in every action. In this sense, our inability to fully compute utility at all of these complex and subjective registers mirrors the struggle of individual agents who make decisions in the world.

B. AN EQUATION FOR CALCULATING HAPPINESS

Bentham's intent was to quantify the overall hedonistically-motivated happiness of individual agents as they make decisions toward the collective good. Equation 1 shows how we incorporate Bentham's variables into the utility calculation in Algorithm 1. The *happiness* (h) of an action's consequences is determined by both immediate and future rewards. The *certainty* (c) and *propinquity* (p) weigh the happiness by its likelihood and delay in its occurrence. The *extent* (e), *intensity* (i), and *duration* (d) are the components of the immediate reward. The *future extent* (e_f), *future intensity* (i_f), and *future duration* (d_f) comprise the estimated future reward and are modified by a pre-defined discount factor (γ).[4]

$$h = cp\left[e(i+d) + \gamma e_f(i_f + d_f)\right] \tag{1}$$

The Bentham circumstances *fecundity* and *purity* are implicitly captured in the future reward. We represent these ideas as magnitudes of happiness rather than as modeled probabilities. In our reading of the source material, Bentham clearly had the right idea that future consequences (which are inherently hard-to-predict) ought to be modified by some discount as one approaches the forecasting horizon. However, the computational language to describe this concept did not exist in his day.

Intuitively, this mathematical structure can be understood as a form of weighted average. An agent calculates the likelihood of an event and multiplies it by its happiness score. In this sense, cp is the weight that scales the happiness score by its probability of occurrence. The remainder of the expression calculates the happiness score in two parts: the current reward and the future reward. The future reward is discounted by γ to reflect the uncertainty of this future happiness.

In each Sugarscape timestep, agents take turns moving, interacting with the environment, and interacting with their neighbors. The turn order is randomized each timestep, and one agent's movement inherently affects subsequent agents. The timesteps then become a Markov decision process. One way to calculate a rational utility measure in a Markov decision process is to use a Bellman equation [76], [77]:

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V^*(s')\right] \tag{2}$$

Here, $V^*(s)$ represents the expected utility of starting in state s and acting optimally. When choosing among actions a from s, the rational choice is to select the action corresponding to the largest utility, or in our case, happiness. Future rewards depend only on s' (the state resulting from action a) and are discounted by $0 \leq \gamma < 1$, since immediate rewards are more valuable than potential future rewards. Equation 1 evaluates happiness in the same way. In particular, $T(s,a,s')$ is cp while the current and future rewards correspond to $e(i+d)$ and $e_f(i_f + d_f)$ respectively.

C. ALTRUISM AND EGOISM

We have presented Jeremy Bentham's hedonic act utilitarianism, but this is not the only ethical theory which can be driven by the quantitative hedonic calculus. There are two additional ethical decision models we implement: *ethical altruism* and *ethical egoism*. Our implementations for both are simple extensions of Algorithm 1. Rather than defining these decision models only in Sugarscape terminology, we first consider the self-driving car (which is rapidly becoming the canonical example of an autonomous decisionmaker [78], [79], [80]) as an explanatory vehicle.

A self-driving car maker includes software to handle all sorts of treacherous road conditions. They are expected to anticipate unlikely situations, but inevitably some self-driving vehicle will be placed in a no-win scenario. The car is essentially presented with a Trolley Problem [78] in which the vehicle cannot avoid disaster, so it is given a choice: it can save its passengers or it can save the bystanders of this incident-in-progress. When the car chooses to save one group it dooms the other to either death or horrible injury.

The vehicle's ethical calculus is informed by the ethical predispositions of the design team and company producing its software. This forces potential passengers and the public to place an immense amount of trust in companies to roll out software which produces rational, reasonable, and reproducible results to moral dilemmas. This requirement of naïve trust is likely an unsatisfactory condition for many given rising calls for ethical guardrails in such systems [81].

All three ethical decision models vary in how they weigh the importance of the individual and everyone else. Ethical egoism is self-interest in the extreme, and an ethical egoist makes choices based solely on what they gain. In the self-driving car example, an egoist car will always save its passengers at the expense of bystanders. Conversely, ethical altruism places no emphasis on the self, and an ethical altruist makes choices solely based on what is to the betterment of everyone but themselves. An altruistic autonomous vehicle will always save the bystanders and doom its passengers. Utilitarianism is a balance between these two extremes and places an equal importance upon everyone. A naïve

[4] We set γ to 0.5, but $0 \leq \gamma < 1$ suffices to provide the necessary decay.

implementation of a utilitarian self-driving car would save whichever group has more members, whether passengers or bystanders. We present further nuance of these three models in our digital terrarium.

D. BENTHAM'S HEDONIC CALCULUS IN SUGARSCAPE

By default, agents in Sugarscape greedily seek to maximize the amount of sugar and spice they collect each timestep. Their goal is to live as long as possible all the while reproducing, engaging in trade, and more. We use Algorithm 1 as a drop-in replacement for each agent's default behavior.

Using the hedonic calculus, the goal remains the same: agents are driven to maximize the collection of sugar and spice. However, the hedonic calculus enables them to consider the outcomes for themselves and those around them. The action which produces the greatest utility is the action which causes the agent and all other nearby agents to experience the greatest amount of communal resource gain (in terms of sugar and spice). This unlocks choices previously unavailable to the agent including (but not limited to) self-sacrifice for the greater good (for example, as an end-of-life decision to unburden society), electing to collect fewer resources to prevent a neighbor from starving, and collecting fewer resources due to having less need than others.

VII. DIGITAL TERRARIUM IN ACTION

We briefly describe the standard configuration of our digital terrarium when generating our experimental results. We calculate agent and societal behavior for 5,000 timesteps for an initial population of 250 agents; for scale, an arbitrary agent can live up to 100 timesteps. Our experimental configuration enables all features described in previous work [7] except for seasons and pollution.[5]

This setup produces a rich and complex social structure mimicking real human societies.[6] Some of the more impactful features we enable are reproduction, trading, lending, inheritance, cultural transmission, combat, and disease. With these features turned on, the simulation is also configured to have three starting *tribes*. A tribe is a group of agents who share a similar set of cultural *tags* used to exert cultural pressure upon others. Agents can engage in combat with members of different tribes but not of the same tribe. Additionally, there are 50 diseases initially spread across the starting agents. We refer to a simulation instance as a *seed*, and we report results as the average performance across 500 different seeds.

We use this uniform configuration to compare four decision models: the naïve greedy model from the source text [6] (which we call *Raw Sugarscape*), an enhanced greedy model which implements ethical egoism to score utility for actions (called *Egoist*), a selfless model which implements ethical altruism to score utility for actions (called *Altruist*), and finally the hedonic act utilitarian approach using the hedonic calculus (called *Utilitarian*).

Egoist, Altruist, and Utilitarian models all use Algorithm 1 and Equation 1. The Egoist model does not consider other agents' preferences when selecting an action. The Altruist model does not consider the deciding agent's preferences when selecting an action. The Utilitarian model equally considers all relevant agents' preferences.

TABLE 3. Survival of decision models across 500 seeds.

Decision Model	Extinct	Worse	Better
Raw Sugarscape	348	5	147
Altruist	344	10	146
Egoist	331	5	164
Utilitarian	0	0	500

In Table 3, we compare the trajectory of societies across 500 seeds after 5,000 timesteps: are they better off than their initial state, worse off (or equal to) their initial state, or dead? Table 3 presents a powerful and succinct summary of our results: utilitarian decisionmaking appears to dominate the other approaches. Even an arguably improved heuristic for greedy decisionmaking (the Egoist model) does not make a meaningful difference at a societal level. Of particular concern is the large proportion of the non-utilitarian societies that go extinct as they appear to be ill-suited to tackle the challenges of adapting to an unknown, hostile environment.

An astute reader may wonder how we determine societal success. This question is particularly relevant since the easily available metrics from Sugarscape inherently reward greedy behavior and seemingly favor the Raw Sugarscape and Egoist decision models. Yet, the cooperative strength of the Utilitarian model far outperforms all three other models despite this implicit advantage for the two greedy models.

We demonstrate the success of the Utilitarian model over the other three across a variety of societal metrics and across multiple experimental setups. These metrics are representative of common-sense measures of societal success and are readily computable in Sugarscape. We propose this set of metrics as necessary pillars for successful societies. Deficiency in any particular metric is indicative of potential societal failure both within Sugarscape and, by analogy, the real world. These metrics are not exhaustive. There may be other equally important measures of societal success to explore in future work.

In Section VII-A, we show the results of societies where all agents adhere to the same model (homogeneous societies). We then demonstrate the effectiveness of utilitarianism in homogeneous societies but alter Algorithm 1 to vary the degree of selfishness in the population. These results on selfishness are described in Sections VII-B–VII-C. We conclude our experimental exploration in Section VII-D with societies using multiple decision models (heterogeneous societies) which acts as a third piece of evidence validating the Utilitarian model's benefits over the other three considered.

[5]These features are implemented but disabled for our experimentation.
[6]Full configuration at: https://github.com/digital-terraria-lab/sugarscape.

Some minor notes on the computational horizon of our simulation are provided in Section VII-E.

A. SOCIO-ETHICALLY HOMOGENEOUS SOCIETIES

We present the result of the same 500 seeds referenced in Table 3 for societies which are entirely composed of Egoist, Altruist, Utilitarian, or Raw Sugarscape adhering agents. The simulation is deterministic. Each seed represents an initial configuration, and deviations between runs of the same seed are attributed solely to differences in decision model applied (as each seed was run four times, once for each model).

We present useful metrics to assess aggregate societal behavior across many seeds. Each metric highlights aspects of successful societies without arguing that these metrics are definitive or complete. When taken in aggregate, we show that Utilitarian agent behavior leads to more promising societies than those with entirely greedy or entirely selfless agents.

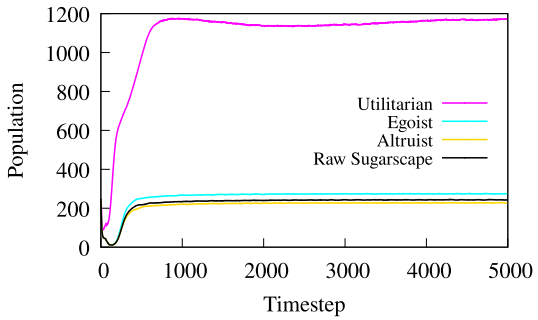


FIGURE 4. Agent population.

One obvious way to evaluate a society is through its population. In Sugarscape, a higher population corresponds to a greater potential to collect resources from an environment. In Figure 4, we provide the mean population across all seeds per timestep. The Utilitarian model quickly achieves a steady state population more than three times greater than the other models. The Egoist model outperforms the Raw Sugarscape model due to its improved heuristic for evaluating greedy actions. The Altruist models only slightly underperforms the Raw Sugarscape model. Altruist agents are prone to self-sacrifice since they care only for the consequences of others and not at all for their own (including premature death).

All four models experience an initial die-off period, a rebounding population recovery, and a sustainable steady population. Based on our observations, this precipitous nosedive can be largely attributed to natural selection, disease, and a predisposition for greedy agents to kill one another for profit (or, in the case of altruism, selfless agents engage in suicide for others to profit). We refer to this as the *murderous period*, a dark age as societies stabilize in hostile environments where a high risk of starvation implies that killing other agents provides vastly more utility than other available actions. Agents who survive the murderous period tend to be young, wealthy, and belonging to a dominant cultural tribe.

Since the die-off rate during the murderous period is so severe in the non-Utilitarian models, many societies simply fail. The average steady state population of these models is dampened by the seeds with 0 population. Regraphing the results for only surviving societies still results in a significantly lower population compared to Utilitarian societies.

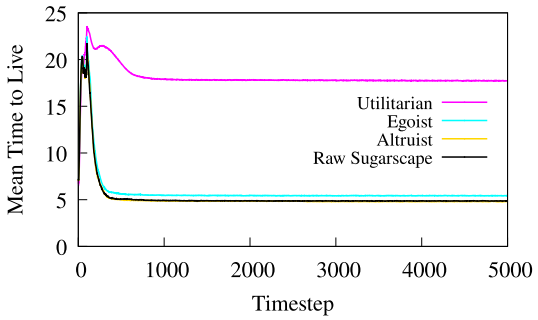


FIGURE 5. Agent mean time to live.

As seen in Figure 5, the Utilitarian model also performs better when considering the agents' mean time to live (TTL). An agent's TTL is the number of timesteps it can survive given its current resources, metabolism, and predetermined maximum age. TTL is a reflection of the quality of an agent's life (i.e. a high TTL corresponds to a longer horizon for a happier life). Combining the observations from Figures 4 and 5, we argue there are more agents, and they are living better with the Utilitarian model than without.

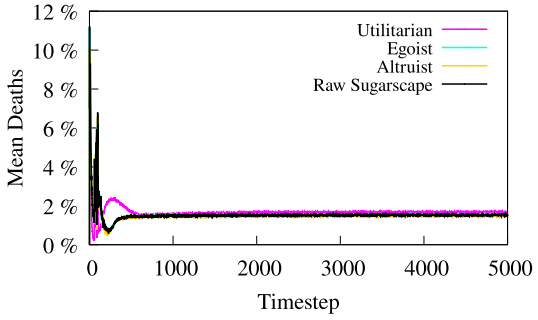


FIGURE 6. Mean deaths per population.

One would expect that overall deaths per timestep would be lower for successful societies. Surprisingly, Figure 6 shows that utilitarian societies have a slightly higher death rate, whereas the other models have very few deaths after the murderous period. We calculate this metric scaled by population, otherwise utilitarian societies (with a relatively high population) would be unduly penalized by the magnitude of their death toll.

Unlike Egoist, Raw Sugarscape, and Altruist agents, Utilitarian agents engage in a democratic process before

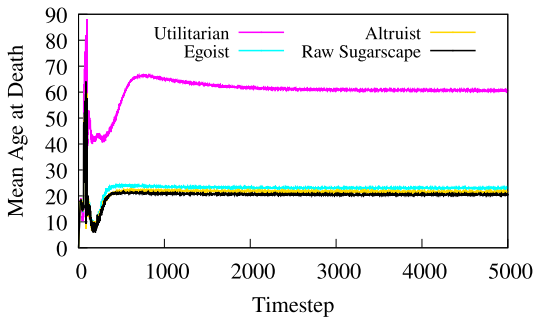


FIGURE 7. Agent mean age at death.

suffering catastrophic consequences for themselves. Doomed agents protest loudly (by expressing their unhappiness in Algorithm 1), however that desire may be drowned out by the overwhelming happiness observed by other affected agents.[7] Furthermore, agents who are otherwise doomed may choose to take end-of-life decisions and pass along wealth for the good of society. These phenomena result in a higher occurrence of death in exchange for greater societal prosperity. As a partial confirmation of its net good for society, Figure 7 shows that Utilitarian agents, while dying more frequently, led far longer and richer (i.e. happier) lives prior to their demise.

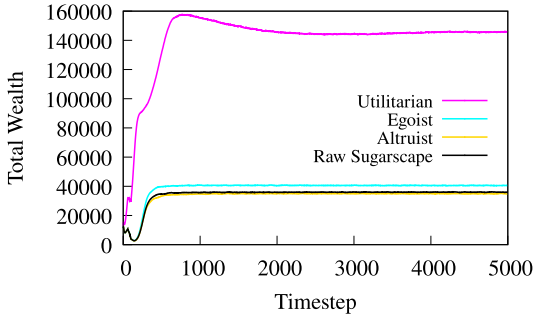


FIGURE 8. Total societal wealth.

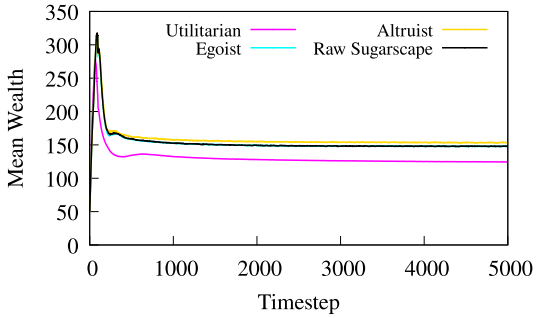


FIGURE 9. Agent mean wealth.

Total wealth in the society is an alluring metric for measuring its success. However, as seen in Figure 8, this metric mirrors total population since more agents will naturally collect more wealth than fewer agents. Yet when we look at an individual agent's average wealth (Figure 9), we see that the other models outperform the Utilitarian model.

After some consideration, one might conclude that since greedy models emphasize individual accumulation over most other priorities, it is not altogether surprising that the data corroborates that logic. What is not readily apparent is that the cost of attaining this larger individual accumulation is a commensurately large number of dead agents (and dead societies) that do not register in the wealth calculation. Accounting for those aspects, the benefit of this behavior is not clear.

For the Altruist model, it is not greed which drives higher mean wealth. Rather, the few agents who are not placed in a position to deferentially commit suicide end up accumulating wealth unabated by others. They bunch up in enclaves of the same tribe where no combat is possible and where self-sacrifice is less likely to occur when alternatives (such as trading or lending) are available.

Among the homogeneous society experiments, it is clear the Utilitarian model is predominate according to the metrics analyzed. The two metrics where Utilitarian societies did worse than the others (mean deaths per timestep and mean agent wealth) are readily qualified by circumstances present in the other societies: in the other models so much of society is violently culled during the murderous period that the remaining agents are either spread far apart or are in enclaves of a single tribe. The lack of competition takes away opportunities for greedy agents to engage in predation and for selfless agents to engage in wasteful self-sacrifice. For all other metrics, Utilitarian societies are far and away the most successful and stable societies.

Algorithm 2 Modified 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
$s \leftarrow d$'s selfishness factor
$utility \leftarrow 0$
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. NOTE ON SELFISHNESS

A significant aspect of Bentham's hedonic act utilitarianism is its emphasis on egalitarian decisionmaking. Everyone is treated equally; only the consequences they experience are

[7] We acknowledge that this unsettling scenario is a key criticism of hedonic act utilitarianism; our goal in this work is to evaluate socio-ethical theories as proposed, not as we wish they could be.

of any importance. This is not true of egoism and altruism. We modify Algorithm 1 to adjust for altruism and egoism by introducing a *selfishness factor* to the decisionmaker. In Algorithm 2, the selfishness factor of the decisionmaker is the weight they provide to their own consequences versus the consequences for others.

An egoist agent has a selfishness factor of 1.0 while an altruist agent's selfishness factor is 0.0. These represent the two extremes of pure selfishness and pure selflessness. The utilitarian is in the middle with a selfishness factor of 0.5. Accordingly, this results in the consequences for all agents being of equal importance (i.e. everyone's happiness score is modified by a factor of 0.5). These variances in selfishness have profound societal impacts.

TABLE 4. Effect of selfishness factor across 500 seeds.

Selfishness	Extinct (%)	Worse (%)	Better (%)
0.00	68.8	5.0	29.2
0.01-0.05	1.5	0.2	98.3
0.06-0.10	1.8	0.2	98.0
0.11-0.15	1.6	0.1	98.3
0.16-0.20	1.4	0.0	98.6
0.21-0.25	1.3	0.1	98.6
0.26-0.30	1.5	0.0	98.5
0.31-0.35	1.3	0.0	98.7
0.36-0.40	1.4	0.1	98.5
0.41-0.45	1.4	0.1	98.5
0.46-0.49	1.4	0.0	98.6
0.50	0.0	0.0	100.0
0.51-0.55	1.2	0.0	98.8
0.56-0.60	1.1	0.0	98.9
0.61-0.65	1.0	0.1	98.9
0.66-0.70	0.9	0.1	99.0
0.71-0.75	1.1	0.1	98.8
0.76-0.80	1.6	0.1	98.7
0.81-0.85	4.6	0.2	95.2
0.86-0.90	19.5	0.9	79.6
0.91-0.95	56.2	1.4	42.4
0.96-0.99	67.2	1.0	31.8
1.00	66.2	1.0	32.8

C. VARYING SELFISHNESS FACTOR

We experiment with varying the selfishness factor of agents to determine just how much and how little greediness is required to result in poor societal outcomes like those shown for the Altruist and Egoist models in Table 3. The Altruist, Egoist, and Raw Sugarscape decision models performed similarly (and poorly) in our previous results across a variety of societal metrics. We concluded that the Utilitarian model is far superior to these other three, but it could be possible that a selfishness factor of 0.5 (as in the Utilitarian model) does not yield optimal outcomes.

We ran our Sugarscape simulation across the same 500 seeds used previously. We varied the selfishness factor for each seed from 0.0 to 1.0 in 0.01 increments (i.e. we explore the full range of 0% to 100% selfishness). Each society is homogeneous as in the previous results in that all agents in a given simulation run have the same selfishness factor and use the modified hedonic calculus in Algorithm 2. Our results reinforce the effectiveness of the Utilitarian decision model, show that the failures present in the Altruist model are easily sidestepped, and demonstrate high societal volatility at high selfishness factor as in the Egoist model.

Table 4 shows the effects of selfishness factor on overall societal success using the same categories from Table 3: extinct, worse, and better. We highlight selfishness factors 0.0 (Altruist), 0.5 (Utilitarian), and 1.0 (Egoist) which match the values from Table 3 exactly. The remaining values are collapsed into groups for readability. The trends are apparent, however. Even the slightest bit of selfishness is enough for largely altruist societies to escape the societal failures of the extreme Altruist model. On the opposite end, there is a tipping point and longer tail in selfishness where society becomes less successful until eventually descending into the poor results of the extreme Egoist model.

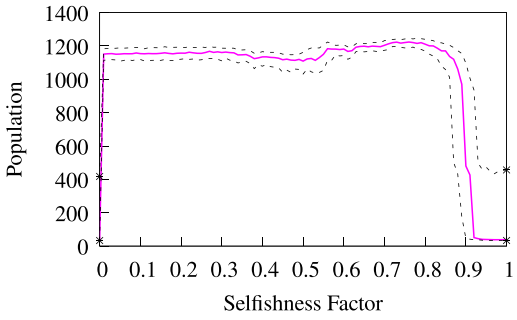


FIGURE 10. Societal population and selfishness.

We provide the same societal metrics discussed previously. Instead of plotting by time, we plot by selfishness factor. Beginning with population, Figure 10 shows the median value of mean population across all 500 seeds per percent of selfishness. The median value is bounded below by the first quartile and above by the third quartile shown as dotted lines.

We see that even the slightest bit of selfishness is enough to pull the Altruist model into success. The pure Altruist model encourages agents to engage in self-sacrifice whenever other agents are even mildly inconvenienced. A common form of self-sacrifice in these scenarios is suicide (by voluntary starvation or by moving to a cell within range of a combatant). A hint of self-interest is enough for an agent to overcome the complaints from others being inconvenienced if it would lead to an unnecessary, early death for the acting agent.

On the other end, there is a longer tail descending toward societal collapse with greater selfishness factors. The more selfish agents become, the more their own voice shouts down the complaints from others. This is especially harmful in cases where an agent behaving selfishly causes the deaths of others (even in cases where the acting agent's survival is not at risk). The more selfish the society, the more volatile its outcomes. The distance between the median and quartiles increases demonstrating that selfish societies have the potential to be marginally successful, but there is also greater risk that society will collapse.

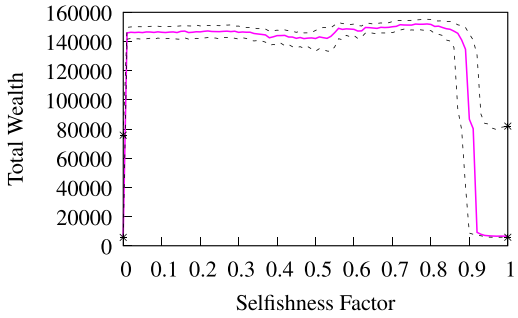


FIGURE 11. Total societal wealth and selfishness.

As previously, societal wealth tracks population. Figure 11 reiterates this. By itself, the result is largely uninteresting.

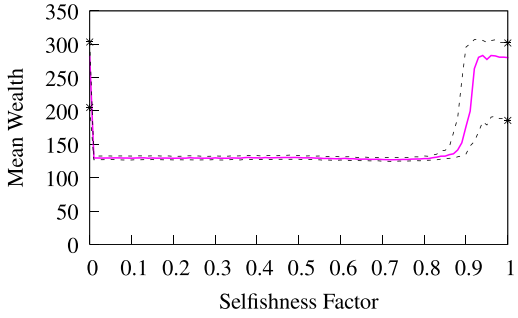


FIGURE 12. Agent mean wealth and selfishness.

However, mean wealth is more informative. Figure 12 shows a similar trend to Figure 9: the more Utilitarian the society, the lower the mean wealth. This taken alone might signal a failure in the Utilitarian model, but our previous results show that these more cooperative agents are simply content living with a lower drive to accumulate wealth. In the more greedy end of selfishness factor, mean wealth is higher at the cost of poor societal outcomes (i.e. a few rich succeed while all others fail) and similarly the Altruist model has higher wealth due to the few agents remaining after rampant self-sacrifice being able to collect resources unobstructed. With less competition, agents can accumulate wealth without inconveniencing others. We again see the increased volatility at the high end of the selfishness factor values.

Figure 13 shows that mean time to live (TTL) has a similar trend to the other metrics. The more selfish the society, the greater its volatility. This volatility represents not only a disparate set of outcomes due to rampant competition but also an increased sensitivity to the initial conditions of the simulation. Some seeds result in an initial placement of agents and diseases as to encourage conflict and contagion. Selfish societies (and self-sacrificing societies in the extreme) cannot adapt and overcome these hurdles in most cases. Additionally, the increased selfishness results in a lower TTL as the wealthy agents prosper while the poorer agents suffer (and die young).

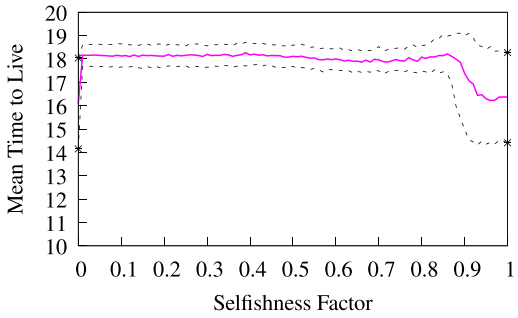


FIGURE 13. Agent mean time to live and selfishness.

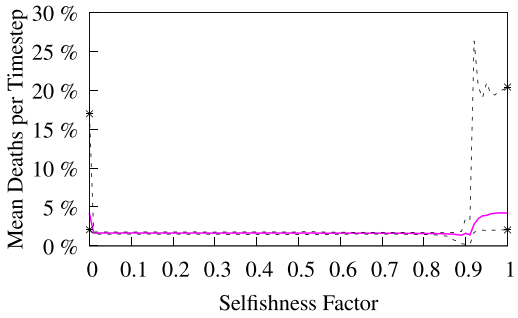


FIGURE 14. Mean deaths per timestep and selfishness.

The intuition that agents in highly selfish societies suffer is in part confirmed in Figure 14 which plots the mean deaths per timestep. Besides the extremes, the overall deaths per timestep is quite low across almost all selfishness factors. However, the trend begins to change around 85% selfishness as the volatility begins to increase. There is a drastic spike after 90% selfishness, indicating incredibly greedy societies risk becoming a meatgrinder where up to a fifth of all agents die per timestep. This kind of nightmarish volatility is indicative of a failed society.

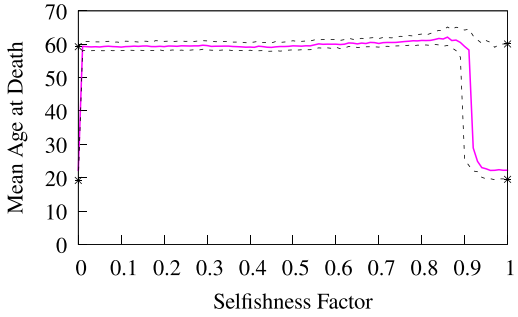


FIGURE 15. Agent mean age at death and selfishness.

The intuition of agents suffering in selfish societies is further confirmed in Figure 15 which plots the mean age at death. At the extremes, agent die young. In high greed societies, the volatility is again high. The median is a paltry

20 timesteps while at best they are as good as the Utilitarian model's performance.

In this series of experiments with a homogeneous population with varied selfishness, we demonstrate the Utilitarian model remains the best performing of the decision models under study. However, the egalitarian approach in the hedonic calculus is not the only way to ensure relative societal success. Any slight amount of self-preservation addresses the pitfalls of the Altruist model. Societies which are mostly altruistic (while not strictly adhering to the Altruist model's 0.0 selfishness) perform on par with Utilitarian societies. This cannot be said for highly greedy societies which fall short of the 1.0 selfishness of the Egoist model. There is a longer tail toward societal collapse than on the Altruist end of the spectrum, and this longer tail also comes with instability. Societal survival becomes a roll of the dice which is an unappealing prospect.

D. SOCIO-ETHICALLY HETEROGENEOUS SOCIETIES

We introduced four different decision models (Raw Sugarscape, Altruist, Egoist, and Utilitarian) and compared their societal outcomes side-by-side. We then compared the performance of societies with different selfishness factors. However, in these experimental setups the entire society of agents behaved in only one way; they adhered to a single decisionmaking principle. Real life is not so simplistic; societies are composed of citizens with differing beliefs. We present experiments with mixed decision models across the same set of seeds and retain the configuration options used previously.

We consider a more realistic, heterogeneous population composed of a mix of Egoist and Utilitarian agents. We ignore the Raw Sugarscape model, since the Egoist model is a more refined heuristic for greedy behavior. Our initial inclusion of the Raw Sugarscape decision model in the homogeneous results is purely for backward readability of *Growing Artificial Societies* [6]. We also ignore the Altruist model since we have established that pure Altruist societies quickly degenerate into rampant self-sacrifice which causes destabilization.

For each seed, we consider different mixes of populations in 1% increments ranging from 0% Utilitarian agents (i.e. all Egoist agents) to 100% Utilitarian agents (i.e. no Egoist agents). One might expect, based on the results of the homogeneous society experiments, that the more utilitarian a society, the more likely it is to succeed. We validate this expectation as a final piece of evidence for our conclusion that the Utilitarian model presents the best outcomes among the four decision models considered.

We begin our analysis the same way as in the single model runs: population. A large population is a strong indicator of a successful society. Figure 16 shows the increase in mean population across the 500 seeds based on the percentage of Utilitarian agents present at the beginning of the simulation. We plot the median of mean populations and one quartile above and below the median shown as dotted lines. Note that

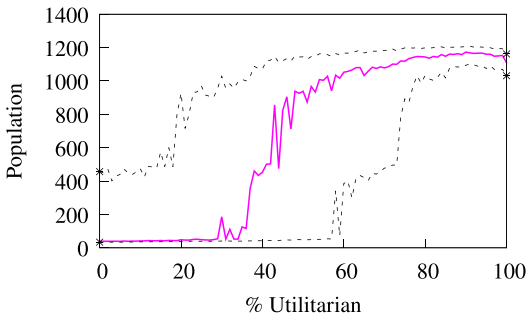


FIGURE 16. Heterogeneous societal population.

the 0% and 100% values match the values for the Egoistic and Utilitarian homogeneous society runs respectively.

It should not be surprising to the astute reader that mean population rises as the relative number of Utilitarian agents increases. After all, the more that agents engage in cooperative decisionmaking, the more effective that society becomes at making community-scale choices. What is more interesting is the progressive reduction in volatility as the percentage of Utilitarian agents increases. In other words, as the ratio of Utilitarian agents increases, a society's population converges more reliably to the median and we see fewer outliers. This provides a second way of seeing the same trend in Figure 10 from the selfishness factor experiments.

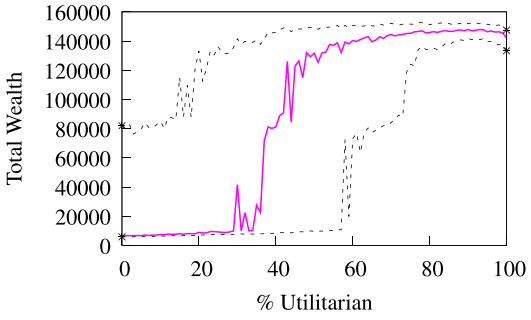


FIGURE 17. Heterogeneous total societal wealth.

As we noted for Figure 8, the total wealth in a society corresponds to the population. As such, Figure 17 is a reiteration of the results of population. The noisy Egoist societies have less total wealth due to fewer agents, and the Utilitarian societies have more wealth and are incredibly stable.

Mean wealth per agent tells us a bit more of the story. Figure 18 shows the tradeoff between Egoist and Utilitarian societies in a way which is obfuscated in Figure 9 for the homogeneous society runs. We see that having a moderate amount of Utilitarian agents (up to roughly 40%) does not meaningfully reduce the average wealth per agent.

After roughly 40% Utilitarian agents in the society, there is an inflection point where the emergent, societal-level behavior switches from that shown in the Egoist

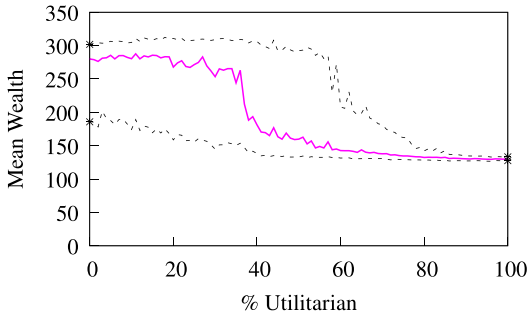


FIGURE 18. Heterogeneous agent mean wealth.

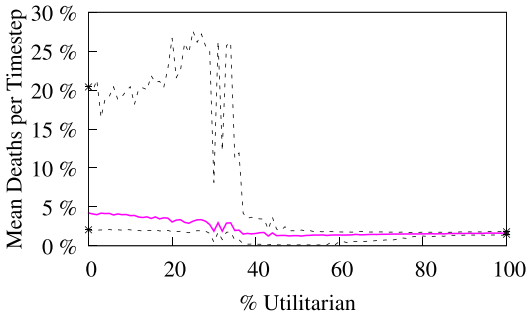


FIGURE 20. Heterogeneous mean deaths per timestep.

homogeneous results to the Utilitarian homogeneous results. For population, societal wealth, and mean agent wealth, this inflection point is quite clear. Society becomes infused with enough cooperative agents to trend societal behavior toward Utilitarian outcomes. As such, volatility begins to reduce (sharply declining around 60% Utilitarian agents). The median value for each of these metrics quickly approaches the value for a homogeneous Utilitarian society.

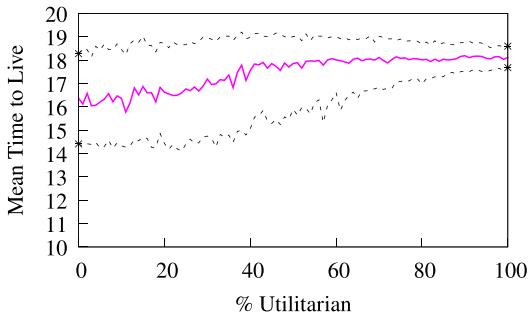


FIGURE 19. Heterogeneous agent mean time to live.

Mean time to live (TTL) presented in Figure 19 shows a slight, steady incline from the full Egoist to full Utilitarian outcome. The lack of sharp changes around 40% Utilitarian population is due to the relatively small difference between the Egoist and Utilitarian models' median performance. The fairly smooth incline of the median value is mirrored by an equally smooth contraction of the volatility in mean TTL. Agents, on the whole, are holding enough resources to continue living longer the more cooperative decisionmaking occurs in the society.

The volatility regarding deaths is most striking among the heterogeneous society results. Figure 20 shows the median value for mean amount of deaths per timestep as a percentage of the population size across all seeds. The median value remains largely static after 40% Utilitarian agents in the society. This is accompanied by a sharp, nearly total elimination, of volatility in the amount of death per timestep.

Majority Egoist societies are subject to either rampant death (due to greedy competition) or very little death (due to the initial murderous period leaving only the rich and powerful remaining). Societal success according to amount of death in these societies is not only hard to predict due to this volatility, but this volatility alone is a marker that these societies are not successful. For an agent living in such a society, their life (and death) becomes a roll of the dice whereas in Utilitarian societies the probability of random death sharply decreases.

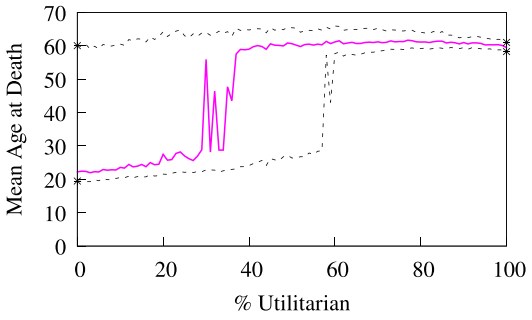


FIGURE 21. Heterogeneous agent mean age at death.

We have shown that agents in Utilitarian societies live longer according to Figure 7. The heterogeneous society result shown in Figure 21 validates this observation and shows a decrease in volatility for mean age at death the more Utilitarian the society. Again, the inflection point at 40% and the drastic reduction of volatility at 60% Utilitarian agents further reinforces there is some critical mass needed for the emergent, societal-level behavior to become cooperative where it was once greedy. Taken together with mean deaths per timestep and mean TTL, the increase in mean age at death shows that the more cooperative (Utilitarian) a society becomes, the more likely it is an agent will know it will die at a relatively old age, without fear of starvation, and in a relatively safe environment without rampant killing.

Table 5 shows the survivability of societies as the percentage of Utilitarian agents increases. Unsurprisingly, the more Utilitarian a society becomes, the more likely the society will end up better off at the end of the simulation from where it started. We note that the previously identified inflection point of 40% Utilitarian agents corresponds with a significant outcome: societies with at least 40% Utilitarian

TABLE 5. Effect of mixing models across 500 seeds.

% Utilitarian	% Extinct	% Worse	% Better
0%	66.2	1.0	32.8
1-5%	66.76	1.92	31.32
6-10%	65.4	1.12	33.48
11-15%	61.92	1.24	36.84
16-20%	58.28	1.2	40.52
21-25%	52.52	1.36	46.12
26-30%	52.52	1.16	46.32
31-35%	50.36	1.64	48.0
36-40%	44.0	1.32	54.68
41-45%	34.6	0.96	64.44
46-50%	30.68	0.64	68.68
51-55%	29.96	0.8	69.24
56-60%	24.68	0.72	74.6
61-65%	21.04	0.64	78.32
66-70%	19.08	0.44	80.48
71-75%	14.28	0.32	85.4
76-80%	10.28	0.48	89.24
81-85%	7.84	0.52	91.64
86-90%	4.4	0.08	95.52
91-95%	1.96	0.2	97.84
96-99%	1.35	0.05	98.6
100%	0.0	0.0	100.0

agents end with a better population more than half the time. The second identified inflection point of 60% also marks a milestone: these societies end up better off three quarters of the time. A fully Utilitarian society all but guarantees a successful society, and there is an incredibly high likelihood of success after reaching the 80% mark which is also reflected across the metrics discussed previously.

E. A NOTE ON THE NUMBER OF TIMESTEPS

An astute reader may be curious whether ending data collection at 5,000 timesteps is sufficient for observing the hypothesized behavior. After all, it is possible some emergent societal behavior, programming error, or configuration misalignment could cause a severe population collapse to occur at timestep 5,001. We demonstrate the robustness of our results in Table 6 which shows a summary for the same 500 seeds used in our experiments for each decision model out to 50,000 timesteps.

Like Table 3, runs of the simulation end in societal extinction, a final population which is worse off than (or equal to) the starting population, or a better population. The values for each category stabilize at 5,000 timesteps and show only minor improvement when running the simulation for the same seeds out to 50,000 timesteps. This incredible stability is convincing evidence that once a society has entered into a steady state, it is almost guaranteed to remain in that steady state. This steady state is representative of a society which has adapted to the ecological carrying capacity of its environment.

There could be a possibility that societal collapse could occur at timestep 50,001. However, this line of inquiry quickly descends into trying to solve a Halting Problem [82], [83]. By demonstrating societal stability out to so many generations, we make the reasonable assumption that no such surprise population death spiral exists after the 5,000 timestep mark. This observation of a steady state (or equilibrium) is validated in the literature for other complex systems [61]. In other words, if a society remains alive at 5,000 timesteps, it is incredibly likely to remain alive.

VIII. SUMMARY OF RESULTS

We have rigorously demonstrated that the Utilitarian decision model leads to more prosperous outcomes in our digital terrarium than the other three models. In all cases investigated, Utilitarian societies always had larger, longer-lived populations with excess societal wealth at their disposal. This pattern is not true for the Raw Sugarscape, Egoist, and Altruist models where more than half of their societies collapsed, and agents often died young and in an impoverished state.

While these other three models do have slightly less agent death and higher average agent wealth, we show that these seemingly positive outcomes have actually come at the cost of great societal pain. The cases where these other models *do* survive nearly succumb to an initial murderous period which Utilitarian societies quickly overcome. The few benefits of living in a greedy, hyper-individualistic society are therefore built on the backs of the many dead and impoverished agents left behind in the wake of the successful few. Likewise, the seemingly pleasant deference in Altruist societies result in rampant and wasteful self-sacrifice which oxymoronically leads to similar societies as the greedy models: exceedingly few remaining agents, likely part of a single dominant tribe, who are free to accumulate wealth while any remaining stragglers live nasty, brutish, and short lives [84].

Our faithful reimplementation of Sugarscape in this work leads us to evaluate societal success based on fundamentally greedy metrics. In this environment, one would presume that locally greedy decision models (such as Raw Sugarscape and Egoist) would far outperform more moderate, forgiving, or considerate decision models. Bentham's hedonic act utilitarianism calculates societal welfare collectively considering all stakeholders' preferences irrespective of the individual decisionmaker's personal biases. However, modern philosophers have long criticized Bentham's work as more closely aligned to capturing economic circumstances than more abstract notions of value even though Bentham clearly defined pleasure more broadly than simply as a metaphor for financial status. In this sense, the results of our experiments are indeed not surprising at all: Bentham's utilitarianism appears to do an excellent job of reaching economic success while incorporating the voices of many agents as a cohesive whole.

To capture the spirit of Bentham's notion of pleasure, we need to find a *quantitative* reward for actions that we know provide more abstract *qualitative* value to an agent and the society as a whole. For example, though teaching usually has some explicit financial gain, it is equally clear that teachers also provide non-monetary, qualitative value for society. In this sense, our work is incomplete; we only capture a fraction of the full utility to which Bentham aspired. There is real work to be done in broadening our digital terrarium's

TABLE 6. Decision model success across various runtimes.

	500 Timesteps			5,000 Timesteps			50,000 Timesteps		
Decision Model	Extinct	Worse	Better	Extinct	Worse	Better	Extinct	Worse	Better
Raw Sugarscape	346	8	146	348	5	147	348	5	147
Altruist	343	13	144	344	10	146	344	8	148
Egoist	329	7	164	331	5	164	331	4	165
Utilitarian	0	0	500	0	0	500	0	0	500

resource collection viewpoint to also incorporate these more holistic notions of value.

IX. REPRODUCIBILITY OF EXPERIMENTATION

The results of this work are eminently reproducible. Our Sugarscape implementation is deterministic and is driven by a user-defined configuration file written in JavaScript Object Notation (JSON).[8] Full software requirements are provided in the `README` of the software repository, but as a highlight the simulation is designed to run on Linux (and likely runs on other mostly POSIX-compliant operating systems) and requires Bash and Python 3. The simulation has a graphical mode which requires a system installation of tk and the tkinter Python module. The results provided in this work were created using the `v2024.1` release. Running the software using a different release will produce similar but not exact results as new features are added and bugs are fixed.

By passing in a configuration as a command line option, one can exactly reproduce the outcome of a given simulation run. We provide the configuration files to reproduce our findings.[9] The configurations are located in four zipped archives:

- `blueprints-homogeneous.zip` contains all the configurations for the initial homogeneous society runs.
- `blueprints-selfishness.zip` contains all the configurations for the varied selfishness factor runs.
- `blueprints-heterogeneous.zip` contains all the configurations for the mixed model runs.
- `blueprints-runtimes.zip` contains all the configurations for the varying runtime note.

When extracted, the configuration files can be fed directly into the simulation software. The simulation takes a single configuration for execution. To run just one iteration of the simulation with one configuration file, provide the following command (using the system's proper Python 3 alias):

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

To run a full set of experiments, for instance the `blueprints-homogeneous` runs, the top-level configuration file `config.json` will need to be modified. The property specifying how many iterations of the simulation will be allowed to run in parallel on the local machine is called `numParallelSimJobs`. Scale this number according to the number of CPU cores or hardware threads available in the machine (minus necessary resources for the operating system). The simulation uses main memory fairly efficiently, so the number of cores is the bottleneck of parallel data collection. The configuration files to use should be placed in the `data` directory of the simulation repository as the execution script uses this as a source to check for existing configurations which do not have a completed log file.

Before performing data collection locally, running `make setup` performs low-effort checking for a local installation of Bash and Python 3. It will then set values in the provided `Makefile` and `config.json` accordingly. The `make data` command will run the data collection, and the `make plots` command will attempt to generate the graphs in this work via the gnuplot graphing software. We note the included plotting script is written with the `blueprints-homogeneous` dataset in mind, and modification will be necessary for generating the other graphs provided in this work. To generate an entire dataset from start to finish and produce graphs, run the following commands:

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

If a distributed data collection approach is preferred, the simulation repository can be readily extended to accommodate. This work utilized the Open Science Grid (OSG) [85] and the HTCondor batch system [86] to perform much of the data collection, and we used the Makeflow scientific workflow management system [87] to structure the jobs submitted to HTCondor. A script can be easily written to create a batch job for each configuration file in the dataset, submit them, and manage their execution. We encourage this approach for more time-efficient reproduction of our results. Note: in practically all distributed execution engines, expect there to be a long tail of total runtime where some straggler jobs take far longer than most to finish. This is most often due to heterogeneity in resources, limited availability to resources at the execution site, or a high load on the system causing all jobs to run slowly on certain execution sites. It may be useful to cancel those straggler jobs when the number of outstanding jobs is small enough to quickly run locally as running `make data` will only (re)run configurations which do not have a matching and *complete* log file in the `data` directory.

X. CONCLUSION AND FUTURE DIRECTIONS

We demonstrate that cooperative societies (composed of utilitarian agents) outperform societies where agents act

[8]Software at: https://github.com/digital-terraria-lab/sugarscape.
[9]Configuration files at: https://github.com/digital-terraria-lab/datasets.

in their own (greedy) interests and where agents engage in wasteful self-sacrifice. In particular, using our digital terrarium, we show that utilitarian societies: are *more than thrice as likely to survive* any given initial landscape, are *nearly six times as populous* as greedy and selfless societies on average, and their *agents live three times longer* lives than their greedy and selfless counterparts.

When varying the degree of selfishness in society, we show that societal survival is practically guaranteed when agents are between 5% and 85% greedy. Such societies attain the performance of the Utilitarian model while the margins outside the range are volatile and prone to failure. When mixing agents with different selfishness factors in the same society, we demonstrate that there are two inflection points. At 40% Utilitarian agents in society, moderate success is achieved. At 60% Utilitarian agents in society, the success approaches that if society was composed entirely of Utilitarian agents. This is further borne out in societal survival. At the 40% mark, societies survive more often than not. At 60%, societies survive three-quarters of the time. This survival rate continues to climb the more Utilitarian agents there are in society.

Our results are quite promising: aside from showcasing how much larger, healthier, and more productive utilitarian societies are than fundamentally greedy or selfless societies, we also make a strong case for the usefulness of our digital terrarium in comparing ethical decision models (in our case: Raw Sugarscape, Egoist, Altruist, and Utilitarian). A key motivation of this project is to provide off-the-shelf reference implementations of popular ethical theories for use in industry applications. We provide a faithful algorithmic representation of Jeremy Bentham's hedonic act utilitarianism, drawn from the original source text [8] and modify it with ethical egoism and ethical altruism. These implemented ethical decision models are simply a starting point for future work.

A. FUTURE WORK

There are many rich and obvious directions to expand this research area. Adding new features to our digital terrarium seems most straightforward; since Sugarscape is a metaphor for real human societies, any model that more closely represents human societies will lead to higher fidelity inferences. Some feature enhancements include: some form of representing social class, distinctions between leaders and followers within tribes, natural disasters, and more thorough climate modeling. These (or other) features would make the simulation more dynamic, complex, and realistic. We particularly invite collaboration with scholars from various fields to add features which accurately represent the contributions of their disciplines in fulfillment of our claim made with Figure 2.

Another avenue would be the expansion of our stable of ethical theories implemented as decision models. Related work [38] demonstrated the beginnings of an implementation of Aristotle's virtue ethics [88] in Sugarscape. A more complete implementation of Aristotle's ethics is appealing both because it is incredibly popular and it is not a consequentialist ethic (like utilitarianism, egoism, and altruism).

There is rich complexity involved with Aristotelian virtue ethics since agents determine their action based on intention rather than calculated consequences. Addressing this complexity would allow us to make direct comparisons between the outcomes of utilitarianism and Aristotelian virtue ethics. This would allow us to perform richer, more realistic experimental comparisons of socio-ethical theories in action.

However, the most tantalizing future direction is a more thorough accounting of all factors in the simulation which lead to prosperity (especially non-economic factors). Jeremy Bentham did not intend for utilitarianism to simply be an economic tool, and he wrote at length about the many different causes of pleasures and pains in life [8]. In our pursuit of this accounting, we need some function to quantify factors in the environment and in agent lives which are seemingly incomparable or inherently qualitative. This is necessary to fulfill Bentham's promise that the hedonic calculus can derive a quantitative scoring of utility from all relevant factors.

Our first inclination is to represent this as a new resource agents can collect. Sugar and spice are collected from the environment, but this third resource (which we will call *nice*) can be collected as a reward for ethical agent behavior and can be traded as transactions of societal goodwill. We must take care not to overly economize this new *nice* currency. This concern along with the creation of a function to quantify non-economic factors are both significant challenges. By overcoming these, we can use *nice* as an agnostic means of quantifying societal success which means it is the best candidate for making direct comparisons between unalike decision models (such as utilitarianism and Aristotelian virtue ethics). Further, this will liberate our agents from acting as *homo economicus* [89], constrained to over-rationalized choices. This allows further decision models patterned after Feminist ethics or Confuscian ethics, among others, to be realized.

B. PARTING THOUGHTS ON THE COMPUTATIONAL TRACTABILITY OF ETHICS

All decisions and behaviors are ethical, in some way. If a decision can be computed, it is no less ethical than if the choice were made by a human person. So, one can say: a computer makes decisions according to its best guesses at some logic of ethics. Taking Sugarscape as originally written, with its greedy agent behavior, we have already coded ethical behavior into an autonomous agent before introducing the richness of our decision layer.

This default, greedy behavior is more or less following the logic of egoistic ethics. Is egoism a form of ethical reasoning? Yes! Is it good ethics? Unlikely. This is the rub in attempting to algorithmify ethics.

Bentham's ethics are one form of ethical reasoning. Can it be computed? Yes, in fact computers can likely bypass some

of the significant weaknesses in Bentham's theory, too, if not the most fundamental flaws. However, this does not avoid the problem that there are *unethical* ethical systems.

The meta-question remaining is whether what it takes to truly live ethically, some definitive and universal ethical framework, is within the reach of human understanding, let alone machine computation. Despite this difficult question, we need ethics. Putting forth no effort to do ethics, nor to make their use possible and transparent for computers, is profoundly irresponsible. To that end, we have followed Bentham's principles as written without claiming his utilitarianism is applicable in all scenarios, for all deployments of artificially intelligent agents. In short: Bentham's utilitarianism does not solve ethics, but it provides a logic with honorable intentions toward outlining ethical behavior in a way which we have demonstrated is computationally tractable.

For Bentham, ethical philosophy begins with a foundational commitment to the alleviation of suffering and maximization of universal benefit. This commitment occurs as a condition for the system he creates to guide decisionmaking. It is Bentham's sense of responsibility for suffering that sets the conditions for his ethical system. It would seem that all ethical systems are predicated on some more foundational philosophy of responsibility. The debate over foundational responsibility rages on, among philosophers, without resolution in sight. Meanwhile, the need for transparency in computational decisionmaking is increasingly pressing.

Instead of taking sides among the philosophical arguments to foundational ethics, we seek an ongoing effort to expand responsibility for suffering and to prevent harm. This is not an abandonment of ethics. It is an unseating of the discipline as some competition of theory (akin to religions with competing truth claims) which currently dominates the field [90].

There is something brilliant unlocked when codifying Bentham's work. Most philosophers teach Bentham with some fondness. He was earnestly trying to do in ethics what folks were doing in the sciences. He was trying to take mythology and religion and hunches and traditions and make them consistent. He loathed the oppression of the poor by the rich, and he was truly revolutionary in suggesting that human lives are all of equal worth.

His system is simplistic, yet this simplicity is precisely what makes his system valuable as the first ethical decision model for our digital terrarium. Bentham's utilitarianism has clear objectives for decisionmaking matching what we expect of machines in predictable decisions. He offers an ethic that is profoundly easier to codify, and therefore makes it easier to be honest and transparent about the decisionmaking process.

Given the advancements in utilitarian philosophy, from John Stuart Mill [91] to Peter Singer [92], it might seem that Bentham was content to play checkers with ethics while others advance to chess. Autonomous agents are only nascently designed to behave according to ethical frameworks. With this in mind, checkers is not such a bad game to learn first.

ACKNOWLEDGMENT

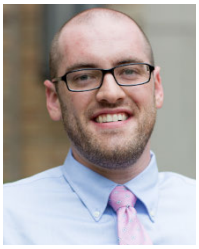
We acknowledge Hadiya Chishti, Colin Hanrahan, Willem Hueffed, Maria Milkowski, Anna Muller, Joshua Palicka, Mariana Shuman, Abi Sipes, and Lucas Vorkoper for their contributions to our software repository. We especially thank Joshua Palicka for his extensivework creating an initial prototype of our Sugarscape reimplementation. We also acknowledge Stuart Glennan for his insightful critiques of our initial approaches to algorithmifying artificial agent ethical decisionmaking. We finally thank Robert Axtell for his enthusiastic guidance regarding our reimplementation of Sugarscape.

REFERENCES

[1] J. McGrath and A. Gupta, "Writing a moral code: Algorithms for ethical reasoning by humans and machines," *Religions*, vol. 9, no. 8, p. 240, Aug. 2018.
[2] D. Castelvecchi, "Can we open the black box of AI?" *Nature*, vol. 538, no. 7623, pp. 20–23, Oct. 2016.
[3] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020.
[4] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, S. Ruggieri, and F. Turini, "Meaningful explanations of black box AI decision systems," in *Proc. 33rd Conf. Artif. Intell. (AAAI), 31st Innov. Appl. Artif. Intell. Conf. (IAAI), 9th Symp. Educ. Adv. Artif. Intell. (AAAI)*, Jul. 2019, pp. 9780–9784.
[5] W. J. von Eschenbach, "Transparency and the black box problem: Why we do not trust AI," *Philosophy Technol.*, vol. 34, no. 4, pp. 1607–1622, Dec. 2021.
[6] J. M. Epstein and R. Axtell, *Growing Artificial Societies: Social Science From the Bottom Up*. Washington, DC, USA: Brookings Institution Press, 1996.
[7] N. Kremer-Herman and A. Gupta, "Replacing sugarscape: A comprehensive, expansive, and transparent reimplementation," in *Proc. Int. Conf. Simul. Tools Techn.* Cham, Switzerland: Springer, 2023, pp. 79–92.
[8] J. Bentham, *An Introduction To the Principles of Morals and Legislation*. London, U.K.: Athlone, 1789.
[9] E. G. Gygax, *Dungeon Masters Guide*, 1st ed., Lake Geneva, WI, USA: Tactical Studies Rules, 1979.
[10] J. von Neumann, *Theory of Self-Reproducing Automata*. Champaign, IL, USA: Univ. Illinois Press, 1966.
[11] C. G. Langton, *Artificial Life: An Overview*. Cambridge, MA, USA: MIT Press, 1995.
[12] T. C. Schelling, *Micromotives and Macrobehavior*. New York, NY, USA: W. W. Norton, 1978.
[13] M. Gardner, "The fantastic combinations of John Conway's new solitaire game 'life' by Martin Gardner," *Sci. Amer.*, vol. 223, pp. 120–123, Jan. 1970.
[14] J. Jara-Ettinger, L. E. Schulz, and J. B. Tenenbaum, "The naive utility calculus as a unified, quantitative framework for action understanding," *Cognit. Psychol.*, vol. 123, Dec. 2020, Art. no. 101334.
[15] A. F. Winfield, K. Michael, J. Pitt, and V. Evers, "Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]," *Proc. IEEE*, vol. 107, no. 3, pp. 509–517, Mar. 2019.
[16] S. T. Segun, "From machine ethics to computational ethics," *AI Soc.*, vol. 36, no. 1, pp. 263–276, Mar. 2021.
[17] T. J. M. Bench-Capon, "Ethical approaches and autonomous systems," *Artif. Intell.*, vol. 281, Apr. 2020, Art. no. 103239.
[18] B. M. McLaren, "Computational models of ethical reasoning: Challenges, initial steps, and future directions," *IEEE Intell. Syst.*, vol. 21, no. 4, pp. 29–37, Jul. 2006.
[19] S. Y. Diallo, F. L. Shults, and W. J. Wildman, "Minding morality: Ethical artificial societies for public policy modeling," *AI Soc.*, vol. 36, no. 1, pp. 49–57, Mar. 2021.
[20] M. Anderson and S. Anderson, "GenEth: A general ethical dilemma analyzer," in *Proc. 28th AAAI Conf. Artif. Intell.*, C. E. Brodley and P. Stone, Eds., Jun. 2014, pp. 253–261.
[21] A. Rand, *The Virtue of Selfishness*. Baltimore, MD, USA: Penguin, 1964.
[22] L. R. Caporael, R. M. Dawes, J. M. Orbell, and A. J. C. van de Kragt, "Selfishness examined: Cooperation in the absence of egoistic incentives," *Behav. Brain Sci.*, vol. 12, no. 4, pp. 683–699, Dec. 1989.

[23] H. Rachlin, "Altruism and selfishness," *Behav. Brain Sci.*, vol. 25, no. 2, pp. 239–250, Apr. 2002.
[24] J. G. Bruhn and J. Lowrey, "The good and bad about greed: How the manifestations of greed can be used to improve organizational and individual behavior and performance," *Consulting Psychol. J., Pract. Res.*, vol. 64, no. 2, pp. 136–150, Jun. 2012.
[25] T. Seuntjens, "The psychology of greed," Ph.D. dissertation, Dept. Social Psychol., Tilburg Univ., Tilburg, The Netherlands, 2016.
[26] E. G. Helzer and E. Rosenzweig, "Examining the role of harm-to-others in lay perceptions of greed," *Organizational Behav. Human Decis. Processes*, vol. 160, pp. 106–114, Sep. 2020.
[27] K. Hoyer, M. Zeelenberg, and S. M. Breugelmans, "Greed: What is it good for?" *Personality Social Psychol. Bull.*, vol. 50, no. 4, pp. 597–612, Apr. 2024.
[28] H. Mintzberg, R. Simons, and K. Basu, "Beyond selfishness," *MIT Sloan Manage. Rev.*, vol. 44, no. 1, pp. 67–74, Oct. 2002.
[29] L. Wang and J. K. Murnighan, "On greed," *Acad. Manage. Ann.*, vol. 5, no. 1, pp. 279–316, 2011.
[30] F. Heylighen, "Evolution, selfishness and cooperation," *J. Ideas*, vol. 2, no. 4, pp. 70–76, 1992.
[31] W. D. Hamilton, "The genetical evolution of social behaviour. I," *J. Theor. Biol.*, vol. 7, no. 1, pp. 1–16, Jul. 1964.
[32] W. D. Hamilton, "The genetical evolution of social behaviour. II," *J. Theor. Biol.*, vol. 7, no. 1, pp. 17–52, Jul. 1964.
[33] R. Dawkins, *The Selfish Gene*. London, U.K.: Oxford Univ. Press, 2016.
[34] W. Williams, *Greed Versus Compassion*. Atlanta, GA, USA: Foundation for Economic Education, 2000.
[35] D. Cassill and A. Watkins, "Mogul games: In defense of inequality as an evolutionary strategy to cope with multiple agents of selection," in *Evolutionary Psychology and Economic Theory*. Leeds, U.K.: Emerald Group Publishing Limited, 2005, pp. 35–59.
[36] J. A. Lasquety-Reyes, "Computer simulations of ethics: The applicability of agent-based modeling for ethical theories," *Eur. J. Formal Sci. Eng.*, vol. 6, no. 2, pp. 75–92, Oct. 2023.
[37] B. Sterner, "Agent-based computer simulation and ethics," *Metascience*, vol. 21, no. 2, pp. 403–407, Mar. 2012.
[38] J. A. Lasquety-Reyes, "Towards computer simulations of virtue ethics," *Open Philosophy*, vol. 2, no. 1, pp. 399–413, Jan. 2019.
[39] B. Smith and C. A. Browne, *Tools and Weapons: The Promise and the Peril of the Digital Age*. Baltimore, MD, USA: Penguin, 2021.
[40] M. Coeckelbergh, *AI Ethics*. Cambridge, MA, USA: MIT Press, 2020.
[41] P. Boddington, *AI Ethics: A Textbook*. Berlin, Germany: Springer, 2023.
[42] L. Floridi, *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. London, U.K.: Oxford Univ. Press, 2023.
[43] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY, USA: St. Martin's Press, 2018.
[44] S. M. Liao, *Ethics of Artificial Intelligence*. London, U.K.: Oxford Univ. Press, 2020.
[45] M. J. Quinn, *Ethics for the Information Age*. Boston, MA, USA: Pearson, 2009.
[46] M. Kearns and A. Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. London, U.K.: Oxford Univ. Press, 2019.
[47] R. Blackman, *Ethical Machines: Your Concise Guide To Totally Unbiased, Transparent, and Respectful AI*. Brighton, MA, USA: Harvard Bus. Press, 2022.
[48] D. Leben, *Ethics for Robots: How To Design a Moral Algorithm*. Evanston, IL, USA: Routledge, 2018.
[49] W. Wallach and C. Allen, *Moral Machines: Teaching Robots Right From Wrong*. London, U.K.: Oxford Univ. Press, 2008.
[50] P. Lin, K. Abney, and G. A. Bekey, *Robot Ethics: the Ethical and Social Implications of Robotics*. Cambridge, MA, USA: MIT Press, 2014.
[51] M. Anderson and S. L. Anderson, *Machine Ethics*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[52] M. Flood, M. Dresher, A. Tucker, and F. Device, "Prisoner's dilemma: Game theory," *Experim. Econ.*, vol. 54, p. 13, Jan. 1950.
[53] R. Axelrod, "Effective choice in the Prisoner's dilemma," *J. Conflict Resolution*, vol. 24, no. 1, pp. 3–25, Mar. 1980.
[54] R. Axelrod, "More effective choice in the Prisoner's dilemma," *J. Conflict Resolution*, vol. 24, no. 3, pp. 379–403, Sep. 1980.
[55] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *Science*, vol. 211, no. 4489, pp. 1390–1396, Mar. 1981.
[56] S. A. West, A. S. Griffin, and A. Gardner, "Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection," *J. Evol. Biol.*, vol. 20, no. 2, pp. 415–432, Mar. 2007.
[57] O. Leimar and J. M. McNamara, "Game theory in biology: 50 years and onwards," *Phil. Trans. Roy. Soc. B, Biol. Sci.*, vol. 378, no. 1876, May 2023, Art. no. 20210509.
[58] J. M. Smith and G. R. Price, "The logic of animal conflict," *Nature*, vol. 246, no. 5427, pp. 15–18, 1973.
[59] J. F. Nash, "Equilibrium points in n-person games," *Proc. Nat. Acad. Sci. USA*, vol. 36, no. 1, pp. 48–49, 1950.
[60] M. A. Nowak, *Evolutionary Dynamics*. Cambridge, MA, USA: Harvard Univ. Press, 2006.
[61] J. Ladyman and K. Wiesner, *What is a Complex System?* New Haven, CT, USA: Yale Univ. Press, 2020.
[62] E. Sangati, F. Sangati, Y.-S. Cheng, and A. Yu-Chan Chang, "Between individual brains and collective behavior: Multi-level emergence in a group formation task," in *Proc. Conf. Artif. Life*, 2023, p. 30.
[63] J. Schossau and A. Hintze, "Towards a theory of mind for artificial intelligence agents," in *Proc. Artif. Life Conf.* Cambridge, MA, USA: MIT Press, 2023, p. 21.
[64] N. B. Ogbo, T. Cimpeanu, A. Di Stefano, and T. A. Han, "Shake on it: The role of commitments and the evolution of coordination in networks of technology firms," in *Proc. Conf. Artif. Life*. Cambridge, MA, USA: MIT Press, 2022, p. 41.
[65] M. Korecki, C. Carissimo, and T. Lund, "ARtificiaL death: Learning from stories of failure," in *Proc. Conf. Artif. Life*. Cambridge, MA, USA: MIT Press, 2023, pp. 1–10.
[66] O. Witkowski and E. Schwitzgebel, "Ethics of artificial life: The moral status of life as it could be," in *Proc. Conf. Artif. Life*. Cambridge, MA, USA: MIT Press, 2022, pp. 1–9.
[67] K. Capek and J. Cejkova, *RUR and the Vision of Artificial Life*. Cambridge, MA, USA: MIT Press, 2024.
[68] J. M. Epstein, *Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science*. Princeton, NJ, USA: Princeton Univ. Press, 2014.
[69] 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*. Berlin, Germany: Springer, 2020, pp. 50–63.
[70] P. Kurakin, "Technoscape: A multi-agent model of all-human global web," in *Proc. 15th Int. Conf. Manage. Large-Scale Syst. Develop. (MLSD)*, Sep. 2022, pp. 1–5.
[71] M. Oremland and R. Laubenbacher, "Using difference equations to find optimal tax structures on the SugarScape," *J. Econ. Interact. Coordination*, vol. 9, no. 2, pp. 233–253, Oct. 2014.
[72] 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 Soc.*, vol. 24, no. 2, pp. 151–164, Sep. 2009.
[73] M. Maleki, N. Nourafza, and S. Setayeshi, "A novel approach for designing a cognitive sugarscape cellular society using an extended moren network," *Intell. Autom. Soft Comput.*, vol. 22, no. 2, pp. 193–201, Apr. 2016.
[74] S. Tisue and U. Wilensky, "Netlogo: Design and implementation of a multi-agent modeling environment," in *Proc. Agent*, 2004, pp. 7–9.
[75] A. Bigbee, C. Cioffi-Revilla, and S. Luke, "Replication of sugarscape using MASON," in *Agent-Based Approaches in Economic and Social Complex Systems IV*. Berlin, Germany: Springer, 2007, pp. 183–190.
[76] R. Bellman, "On the theory of dynamic programming," *Proc. Nat. Acad. Sci.*, vol. 38, no. 8, pp. 716–719, Aug. 1952.
[77] R. Bellman, "A Markovian decision process," *J. Math. Mech.*, vol. 6, pp. 679–684, Jan. 1957.
[78] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, Nov. 2018.
[79] S. Karnouskos, "Self-driving car acceptance and the role of ethics," *IEEE Trans. Eng. Manag.*, vol. 67, no. 2, pp. 252–265, May 2020.
[80] A. E. Jaques, "Why the moral machine is a monster," *Univ. Miami School Law*, vol. 10, pp. 1–10, Apr. 2019.
[81] A. Boyle, J. Thorne, N. Bek, T. Soper, L. Stiffler, K. Schlosser, R. Yonck, T. Wilde, T. Bishop, C. Schubert, K. Gill, N. Graham, and F. Catalano, "Bot or not," GeekWire, Seattle, WA, USA, 2024.
[82] A. M. Turing, "On computable numbers, with an application to the entscheidungsproblem," *J. Math.*, vol. 58, pp. 345–363, Jan. 1936.
[83] M. Davis, *Computability & Unsolvability*. New York, NY, USA: McGraw-Hill, 1958.
[84] T. Hobbes, *Leviathan or The Matter, Forme and Power of a Commonwealth Ecclesiasticall and Civil*. London, U.K.: Printed for A. Crooke, 1651.

[85] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, and F. Wurthwein, "The open science grid," *J. Phys., Conf.*, vol. 78, Jul. 2007, Art. no. 012057.
[86] D. Thain, T. Tannenbaum, and M. Livny, "Condor and the grid," in *Grid Computing: Making Global Infrastructure a Reality*. Madison, WI, USA: University of Wisconsin–Madison, Department of Computer Sciences, 2003, pp. 299–335.
[87] M. Albrecht, P. Donnelly, P. Bui, and D. Thain, "Makeflow: A portable abstraction for data intensive computing on clusters, clouds, and grids," in *Proc. 1st ACM SIGMOD Workshop Scalable Workflow Execution Engines Technol.*, May 2012, pp. 1–13.
[88] J. Sachs, *Nicomachean Ethics*. Indianapolis, IN, USA: Hackett, 2011.
[89] D. A. Urbina and A. Ruiz-Villaverde, "A critical review of homo economicus from five approaches," *Amer. J. Econ. Sociol.*, vol. 78, no. 1, pp. 63–93, Jan. 2019.
[90] E. R. Severson, *Before Ethics*. Dubuque, IA, USA: Kendall Hunt, 2022.
[91] J. S. Mill, "Utilitarianism," in *Seven Masterpieces of Philosophy*. Evanston, IL, USA: Routledge, 2016, pp. 329–375.
[92] K. de Lazari-Radek and P. Singer, *Utilitarianism: A Very Short Introduction*. London, U.K.: Oxford Univ. Press, 2017.



NATHANIEL KREMER-HERMAN (Member, IEEE) received the B.A. degree in computer science with minors in philosophy and sociology from Hanover College, Hanover, IN, USA, in 2015, and the M.S. and Ph.D. degrees in computer science and engineering from the University of Notre Dame, Notre Dame, IN, USA, in 2019 and 2021, respectively.

From 2015 to 2020, he was a Research Assistant with the Cooperative Computing Laboratory, University of Notre Dame. From 2016 to 2017, he was a Teaching Assistant with the University of Notre Dame, where he was a Teaching Fellow, in 2018, and finally an Instructor, from 2019 to 2020. From 2020 to 2022, he was an Assistant Professor with the Department of Computer Science and the Department of Engineering, Hanover College. Since 2022, he has been an Assistant Professor with the Department of Computer Science, Seattle University, Seattle, WA, USA.

Dr. Kremer-Herman is a member of the Association for Computing Machinery (ACM) and their Special Interest Group on Computer Science Education (SIGCSE). He is also a member of the Consortium for Computing Sciences in Colleges (CCSC). He is a member of the IEEE Systems, Man, and Cybernetics Society and the IEEE Systems Council. He has an Erdős number of 4.

ANKUR GUPTA (Member, IEEE) received the B.S. degree in computer science and in mathematics and the M.S. degree in computer science from The University of Texas at Dallas (UT Dallas), Richardson, TX, USA, in 2000, and the Ph.D. degree in computer science from Duke University, Durham, NC, USA, in 2007.

He was a Lecturer with the Computer Science and Software Engineering Department, Butler University, Indianapolis, IN, USA, until 2008, where he was an Assistant Professor, until 2013, and an Associate Professor, until 2021. He is currently a Professor and the Chair of the Computer Science and Software Engineering Department, Butler University. He is an internationally recognized expert in algorithm design and analysis, with a particular emphasis on text (and pattern) matching, data compression, and big data applications. In particular, he developed the wavelet tree, a fundamental data structure that has revolutionized myriad areas in computer science. His most recent work studies the concepts of artificial wisdom (supported by a prior grant from the John Templeton Foundation) and artificial ethics.

Dr. Gupta is a Lifetime Member of the Association for Computing Machinery (ACM). He is also a member of the IEEE Systems, Man, and Cybernetics Society. He has an Erdős number of 3.

ERIC R. SEVERSON received the B.A. degree in religious studies from Northwest Nazarene University, Nampa, ID, USA, in 1996, the M.Div. degree from Nazarene Theological Seminary, Kansas City, MO, USA, in 2000, and the Ph.D. degree in religious studies from Boston University, Boston, MA, USA, in 2010.

He was an Assistant Professor of philosophy and religion with the Eastern Nazarene College, Quincy, MA, USA, from 2004 to 2009. He was an Associate Professor of philosophy with the Eastern Nazarene College, from 2009 to 2013. He was the Director of the Center for Responsibility and Justice, Eastern Nazarene College, from 2009 to 2013. He was the Chair of the Division of General Education, Eastern Nazarene College, from 2011 to 2013. Since 2016, he has been an Associate Teaching Professor of philosophy with Seattle University, Seattle, WA, USA. He is the author of *Before Ethics*, *Levinas's Philosophy of Time*, and *Scandalous Obligation*.

Dr. Severson is a fellow of the Psychology and the Other Institute. He is a Charter Member and the past President of the Wesleyan Philosophical Society. He looks forward to having an Erdős number of 4.