INSTITUT FÜR INFORMATIK
der Ludwig-Maximilians-Universität München

Overview of Computational Ethics

Ardit Mazreku

Bachelorarbeit

Betreuer Prof. Dr. François Bry
Mentorin Nemi Pelgrom

Abgabe am 14.04.2022

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst habe und keine anderen als die angegebenen Hilfsmittel verwendet habe.

München, den 14.04.2022 Ardit Mazreku, Matr.: 11584506

Abstract

Computational ethics is the application of computational approaches to ethical questions or to the investigation of the evolutionary emergence of ethical behavior. Computers and simulation are being used to solve ethical issues.

This bachelor thesis gives an overview of the current state of the research in computational ethics, by focusing on what ethical issues are investigated in this field using simulation, by describing and motivating the state of the current research, and by describing the different methodologies used in the several ethical questions being addressed. The different modeling approaches covered are agent-based modeling, and the use of game theory and logic programming related to ethics. The first mainly investigates specific behaviors in groups of individuals and how different conditions affect them. The second explores strategies followed by individuals for decision making. And the last examines how to build frameworks for ethical judgments and model morality. Next it discusses the different properties of these approaches and examines the significance of modeling and simulation in ethics.

Zusammenfassung

Computational Ethics ist die Anwendung computergestützter Ansätze auf ethische Fragen oder die Untersuchung der evolutionären Entstehung ethischen Verhaltens. Computer und Simulationen werden eingesetzt, um ethische Fragen zu lösen.

Diese Bachelorarbeit gibt einen Überblick über den aktuellen Stand der Forschung im Bereich der Computational Ethics, indem sie sich darauf konzentriert, welche ethischen Fragen in diesem Bereich mithilfe von Simulationen untersucht werden, indem sie den Stand der aktuellen Forschung beschreibt und motiviert, und indem sie die verschiedenen Methoden beschreibt, die bei den verschiedenen ethischen Fragen, die behandelt werden, verwendet werden. Die verschiedenen Modellierungsansätze, die behandelt werden, sind die Agent-Based modeling, die Verwendung der Game Theory und logische Programmierung im Zusammenhang mit der Ethik. Der erste Ansatz untersucht vor allem spezifische Verhaltensweisen in Gruppen von Individuen und wie verschiedene Bedingungen diese beeinflussen. Im zweiten werden die Strategien untersucht, die die Individuen bei der Entscheidungsfindung anwenden. Im letzten, wird untersucht, wie man einen Rahmen für ethische Urteile schaffen kann und wie Moral modelliert werden kann. Anschließend werden die verschiedenen Eigenschaften dieser Ansätze erörtert und die Bedeutung von Modellierung und Simulation in der Ethik untersucht.

Acknowledgments

I would like to thank Professor Bry for giving me the opportunity to work on a thesis that combines computer science and philosophy, and for providing support and encouragement throughout this project.

I would also like to thank Nemi Pelgrom for providing support, guidance and encouragement throughout this project.

Contents

CHAPTER 1

Introduction

Computational ethics is the application of computational approaches to ethical questions or to the investigation of the evolutionary emergence of ethical behavior. It consists of modeling ethical systems with the intent of observing their dynamics and exploring the relationship between individual ethical actions and their contributions to the evolution of a large scale emergent ethic. Thus, there will be no no focus on topics such as the philosophy of computation, AI, information or technology, nor the social impact of computer use or computer ethics.

Computational ethics allows us to experiment with, and test, social ethical theories, to facilitate quantitative research in ethics, to test ethical frameworks, to analyze individual ethical principles and the moral interrelationships that may arise between an individual and its group. Furthermore, it gives the opportunity to explore the consequences of this interrelationship in a uniquely structured environment. Computational ethics addresses the individual behavioral manifestations of an ethic, as well as the emerging social consequences, to which individual actions contribute. Additionally, computational ethics provides a methodological framework for studying the nature of computational worlds in which certain ethical principles prevail and others in which these same principles do not. [76]

More practically, computational ethics can help to investigate questions such as:

- how can experiments for ethical theories be created?
- is it possible to model theories in normative ethics?
- can virtue ethics be used to help individuals make decisions on how to behave?
- are there quantifiable benefits to altruism?
- can there be a scenario where unethical actions can be considered altruistic and have any ethical value?
- how do values and norms affect the way individuals act?
- under which conditions are clearly unethical behaviors repeated?
- how can meta-ethical problems be approached with computational methods?

The focus of this thesis is to investigate which methodologies have so far been used in answering such questions, with the intention of giving an overview of the current state

of research in computational ethics. Furthermore, this thesis also aims to find out which specific computational approaches are chosen to solve these problems, and how they are implemented.

The computational approach this thesis mostly focuses on is agent-based modeling as it provides a clear correspondence to the individual and group interrelationship. Alternative approaches such as game theory and logic programming will also be considered to see how these differ from agent-based modeling.

This thesis consists of five more chapters, structured as follows. The next chapter goes more in depth into defining computational ethics and motivating the current research, it also contains definitions of concepts discussed in the thesis. The following chapter defines more precisely agent-based modeling as considered in this thesis and gives a detailed description of a representative example, with model and simulations descriptions. Next, Chapter 4, lists several models relating to ethics, divided into the categories models of ethical issues with agent-based modeling, with game theory, and with logic programming. Each model description contains a short description of the simulation and its results followed by the model description and concluded by a sentence describing the addressed ethical issue. Next, follows a discussion about the work presented, considerations and counter arguments about the application of computational methodologies to ethics and alternative approaches. The final chapter contains the closing remarks.

CHAPTER 2

Computational Ethics

This chapter defines what computational research is and motivates research in this field. It also provides some additional information about modeling and simulation, which are an integral part of computational ethics.

2.1 Definition

Theoretical studies in evolutionary ethics and experiments with artificial life suggest ways in which ethical behavior may emerge in autonomous agents. [1]. Moor defines computational ethics as the subject of inquiry focused on actualizing how artificial intelligence systems might make ethical decisions [62]. Computational ethics is concerned with the computational complexities required to build intelligent systems to make ethical decisions, as well as what might constitute the computational threshold to consider these systems as ethical artificial agents [63, p. 222]. The aim of computational ethics is to strip ethics of complexities and making it computable [1]. In a very recent article, Portmann and D'Onofrio explain what computational ethics is by mentioning that it bridges concepts of traditional ethics into computer artifacts. [73, p. 5]. As mentioned in the introduction and building on the definitions from the literature presented above, the following definition of computational ethics is the one this thesis is concerned with:

Definition 2.1.1 (Computational Ethics)**.** Computational ethics is the application of computational methods to ethics, with the aim of tackling ethical issues.

An example of the scientific work in this research field is Körner's master thesis "On The Origin Of Altruism: An Agent-Based Social Evolutionary Simulation" 4.1.1. It investigates the dynamics of the evolution of altruism using agent-based modeling. It does so by designing agents with specific traits relating to altruistic actions and comparing them to a control group, that does not possess these traits, and then evaluating the results.[52]

2.2 Background

2.2.1 Ethics

Ethics, described very simply, is that part of philosophy that investigates how people should act. The word "ethics" is derived from the Greek word *êthikos*, which means manners and customs. "Moral" is derived from the Latin word *moralis*, which also means manners and customs. Thus, the two are generally used interchangeably. Ethics or ethical philosophy is a branch of philosophy that studies the principles of what is right or wrong in human manners or human behavior [70] and what is good and evil, virtue and vice, justice and crime. It involves systematizing, defending, and recommending such concepts and seeks to resolve questions of human morality by defining these concepts. Three major areas of study within ethics are [101]:

- Normative ethics: concerning the practical means of determining a moral course of action. It examines questions such as: How do we tell if something is good or bad? Is it good if it maximizes happiness? If it comes from good intentions? If it fulfills the purpose of humans? If you could make it a universal law? If it fulfills certain rights and duties?
- Applied ethics: concerning what a person is obligated (or permitted) to do in a specific situation or a particular domain of action. It examines questions such as: Is abortion moral? Do corporations have responsibilities? How should we allocate scarce resources? When is lying okay? Is file sharing theft?
- Meta-ethics: concerning the theoretical meaning and reference of moral propositions, and how their truth values, if any, can be determined. It examines questions such as: What do moral sentences mean? What is rightness and wrongness? Does moral value exist in the universe?

2.2.2 Modeling

Regarding the motivation for modeling, Epstein answers with the following. Models come from assumptions, which are laid out in detail, so the exact requirements can be studied. These assumptions, produce specific results, and when the assumptions are altered, different results are produced. These results can be replicated by others or calibrated to historical cases, if there are data, and can be tested against current data to the extent that they exist. All this can then be incorporated in the best domain, e.g. ethics, expertise in a rigorous way. By revealing trade-offs, uncertainties, and sensitivities, models can discipline the dialogue about options and make unavoidable judgments more considered. Some of the main reasons to model are to: explain what is being modeled; guide data collection; illuminate core dynamics; suggest dynamical analogies; discover new questions; promote a scientific habit of mind; bound outcomes to plausible ranges; illuminate core uncertainties; offer crisis options in near-real time; demonstrate trade-offs; suggest efficiencies; challenge the robustness of prevailing theory through perturbations; expose prevailing wisdom as incompatible with available data; train practitioners; discipline the policy dialogue; educate the general public; reveal the apparently simple to be complex and vice versa [41]. Building on those reasons, the use of computers to simulate and study complex systems using mathematics, physics and computer science that results in computational modeling can be motivated [67].

2.2.3 Simulation

Simulation modeling and analysis is the process of creating and experimenting with a computerized mathematical model of a physical system [29], imitating the behavior of a real-world process or system over time [15]. Its benefits are experimentation in compressed time, reduced analytic requirements, and easily demonstrated models. Whereas, its limitations are that simulation cannot give accurate results when the input data are inaccurate, provide easy answers to complex problems, and solve problems by itself [29].

Computer simulations are computer process that mimic features of a target physical process, such that a common dynamical theory is capable of describing both the simulation and its target process. For practical and theoretical reasons, a simulation is strictly simpler than its target. There is a homomorphism from the target process to the simulation and no isomorphism. [56] From here on, computer simulations will simply be referred to as simulations.

2.3 Motivation

Simulation allow us to observe the evolution of the society's strategic equilibrium via the individual interactions experienced by the agents in the system. This strategic equilibrium may be mapped to the society's ultimate social morality in order to give insight into the effects of various ethical principles within a society on its overall moral landscape [76].

The interactions of agents with other agents, or the environment, can lead to different results depending on the assignment of specific conditions and values. They can also lead to unforeseen or surprising emergent behaviors, where a complex property at the macro system level is produced that is not encoded at the individual agent level [13, p. 4] [104, p. 29]. The possibility of running such computer simulations over and over with different variables make them function like digital laboratories where one can perform experiments and test hypotheses [42, p. 4] [43, p. 14]. They are particularly attractive for social scientists because many social experiments cannot be practically or ethically carried out in the real world. [53] For the same reason they can be attractive for ethicists.

Segun also mentions that computational ethicists are concerned with developing the decision making architecture of artificial intelligence systems, so they can make ethical decisions [82]. Computational ethics also raises practical questions on the plausibility and the tractability of ethical principles as they apply to artificial intelligence systems [26]. In cases where it is largely evident that an ethical principle is not calculable or possess non-procedural features, computational ethicists are tasked with designing an analytic framework to validate the usefulness of these principles. In other words, with knowledge representation and reasoning, computational ethicists develop the semantic and syntactic functions required to represent these abstract ethical principles in forms that are computable for artificial intelligence systems [55]. [82]

The aim of computational ethics is also to apply critical and practical models to ethical principles by maintaining logical consistency. Its main concerns are defining practical steps to codifying ethics [82].

Computational ethics claims that at least some components of good ethical decision making are computational in nature. Hence, some combination of algorithms, equations, heuristics, rules, or networks may eventually serve as a basis, or at least part of a basis, for good ethical decision making. [62] Examining ethical theory from a computational perspective gives a fresh, critical outlook on a difficult subject. Considering how to mode ethical decision making forces our hidden assumptions about ethics to the surface for scrutiny. If

we are really committed to a moral point of view, really committed, then precise thinking about ethical theory is crucial in order to make the best decisions. [62]

Objections about computational ethics are based on the fact that the only model for a moral agent is a human being, who is associated responsibility for actions with accompanying praise and blame. Which is not applicable to artificial moral agents, as it does not make sense to praise or blame them. However, they can be changed so that they will not decide to act improperly in the future. One way to use computers in ethical decision making is to rely on them to abstract patterns of information from an otherwise indecipherable blur of data. The division of labor between humans and computers with regard to ethical decision making is a distinction between cognition and computation. [62]

There are different approaches to investigating ethical questions computationally. One is to use ABM to represent individuals in a society and simulate different behaviors and relations which allows the investigation of what might happen under different circumstances. These ABM themselves can be structured in different ways, and the main difference might lie in the agents' decision making, these might be constructed based on empirical data as if-else statement or as a probability (where the probability distribution is based on the empirical data), as a optimization function, etc. Another approach is to use game theory and, therefore analyze and predict how agents behave in strategic situations [30]. A different approach still is to use logic programming to define formally actions that might be taken, or to create an analyzer which can assist in practical ethical questions, such as "Should a healthcare professional try to change a patients mind, who rejects a helpful treatment or accept the patient's decision?" [8]

In their book *Evolving Ethics: The New Science of Good and Evil*, Mascaro, Korb, Nicholson and Woodberry introduce agent-based modeling to experimental philosophy and show that agent-based computer simulation is a viable way of studying ethics. They argue that simulations of ethical decisions are epistemologically equivalent to experiments with human subjects, without having to unethically experiment on human subjects [92].

Next, follows a more detailed description of agent-based modeling, and of a representative example of a model and some simulations applying it to different ethical issues.

CHAPTER 3

Agent-based Modeling

This chapter introduces the definition and motivation for agent-based modeling and how it relates to computational ethics. Followed by an detailed description of a representative example of an agent-based model of an ethical theory, with the goal of showing an accurate example of the research in computational ethics and using this as a basis for the comparison with other agent-based models and their relation to ethics.

3.1 Definition and Motivation

Agent-based modeling, short ABM, is a methodology where a collection of autonomous decision-making entities called agents is used to build formal models of real-world systems [24] [28].

It establishes a direct correspondence between the real world individual units in the target system to be modeled, and the parts of the model that represent these units, that is, the agents interactions of individual units in the target system and the interactions of corresponding agents in the model. Therefore, ABMs are used to simulate the actions and interactions of autonomous agents in order to understand the behavior of a system and what governs its outcomes. The main benefit of using this method is that it captures emergent phenomena, that is, phenomena resulting from the interactions of individual entities. An investigation of large scale patterns, macro patterns, which evolved from the interactions of numerous interacting micro-agents and cannot be reduced to the system's parts. ABMs should be implemented if the individual behavior to be modeled is nonlinear and can be characterized by if-then rules, if the individuals to be modeled exhibit memory capacity, if their decisions depend on path taken, and if they exhibit non-markovian behavior, that is, the probability of undertaking a set of actions is variable. [28] By this description, ABMs allow the simulation of individuals interacting with each other and undertaking actions which might affect themselves, other individuals or have a big impact on the group if repeated for a relatively long period of time. These properties of ABM allow the overcoming of limits and difficulty of real world exploration of ethics, by simulating practical research in this field.

The next section describes an example of an ABM used to tackle different ethical issues. The first part describes the model specification in sufficient detail to give an idea of the state of the research using ABMs in ethics, and generally enough that is applicable to the simu-

lation of different ethical issues. The model description does not contain information about specific parameters, values and formulas used, as these differ slightly for each simulation described in the second part of this section.

3.2 Representative Example

The following model and simulations are described in detail by Mascaro, Korb, and Nicholson in their book *Evolving Ethics: The New Science of Good and Evil* [56]. The version described next, or modified versions coming from it, are used in this book for different simulations related to evolutionary psychology and ethics.

3.2.1 Model

The following paragraphs describe the environment of the model, how the passage of time is represented, the entities that populate the environment, that is, food and agents, and how the latter can act and make decisions. This description is taken from [56, p. 86-95]. The subsection is concluded with a summary of the model description.

The world of the model, as in most ABMs, is a two-dimensional grid of *nxn* cells, where the size varies based on the experiment. Each cell can be occupied by an unlimited number of entities or be limited to maximum one. The world can either have bounds at the edges or opposing edges can be connected (torus shaped).

Time in the simulation passes in cycles (also called steps). A cycle is a randomly ordered loop through all agents in the simulation, in which each gets the chance to act. A cycle can consist of two passes: a first one to let agents observe before choosing an action; and a second one to let them to act. A period of cycles during which the system collects statistics is an epoch.

Food is one of the entities that can occupy cells in the world. It exists to fuel agents and have the following properties: it can occupy a cell; it can be consumed, which removes the food from the cell; it can have a finite lifetime, even if not eaten; each piece of food has an associated health that is used to boost an agent's health when the agent eats it; it is generated by the system using the food distribution function (fdf) mapping the current time to an amount of food, which can be a constant function or sinusoidal function, to describe seasonal changes.

Agents are the other entity that can occupy cells. They have divergent properties and behaviors depending on the experiment. They can be created by sexual or asexual reproduction. They possess an *age*, which is the number of cycles passed since birth, and they can observe their own age. Gaussian simulation parameters are used to give to each agent a maximum age at birth. Agents might die before they reach their maximum age, if their health is exhausted, by accident or by committing suicide. Once dead, the agent is removed from the board. The agents' *health* is represented by a numerical value. They start with an initial health value, which they get as initial investments of health by each parent (this can be fixed or variable). At the beginning of the simulation, each agents' initial value is several times higher than the average parental investment. The health value changes during the simulation according to actions performed. For example, eating food increases agent health by food health value minus cost of eating food - if the food is poisonous the agent's

health is decreased - whereas resting increases the agent's health, although it is a much smaller increment than eating.

The *actions* available to the agent occur within its neighborhood (cell occupied by the agent and surrounding eight cells, in some experiments it is larger), and may be condition on the agent's observations. If an action can be performed, it will be performed. The selection of an action is done by the agent's evolvable decision function, mapping observations of the environment to actions (the output is clipped and scaled as the interval $[0, 1]$). Each action and their outcomes have a positive or negative utility, whose value is established by a fixed or evolving utility function. This utility function calculates the value of an action on its individual, consequential merits, as described by act utilitarianism. The utilitarian value function $v(a)$ of action a is:

$$v(a) = \sum_i \sum_j u_i(o_j) p(o_j \mid a)$$

where u_i is the utility function of agent i, o_j is a set of possible outcomes of actions a, and $p(o_j \mid a)$ is the probability of the outcome o_j given a. Agents can *observe* different properties about themselves (health, age, sex, gestation status) or the environment (local or global population density, local or global food density, and food availability) The basic actions they can perform are eating, moving, resting and reproducing. Whereas some experiments have additional actions (suicide, rape, abortion). Agents' *movement* (migration, transmission) consists in changing from the current position to one of the neighboring cells. Depending on the experiment, this can happen in a single action, to a random neighboring cell; or divided into two actions: first turning, then walking. If no movement is possible, the agent simply rests.

Simulations contain asexual, sexual agents, or both. They can have genderless sex (i.e., any two individuals within a species can mate) or they can have genders - female sex is the one that gestates and/or the one that invests more in the offspring. Sexual *reproduction* can have a gestation period prior to birth; or birth is immediate. Upon birth, the offspring is placed somewhere within the neighborhood of its parent, or, if kin selection is turned off, may be placed anywhere. *Mating* Can be non-consensual and result in a big negative utility. Choosing mates consists in making a request which can be turned down. Production of offspring will fail if: the agent lacks sufficient health or maturity; there is no viable partner available; there is no available space for the offspring. In the absence of consensual mating there is no negative utility.

Agent Genotypes condition the decision making of the agents and over time they may become better adapted to their environments. This is mainly reflected in improved decision functions encoded in their chromosomes, which are either sets of production rules, or decision trees. Both types have their own structure, usage, crossover method and mutation method. In addition, agent genomes may also possess variables representing mutation rates, parental investments and an age of expiration. They may also hold bit signatures for an agent's immune system or mate compatibility and a disease's infectiousness or virulence

To summarize, agents observe their environment and use these observations to condition their actions using on an evolvable decision function. The system collects various demographic statistics, and statistics on action rates and on genetic properties. This generic simulation design is both simple and extensible and can be adapted easily to new problems.

The evolutionary approach is intended to help to better understand the evolutionary circumstances in which a kind of ethical behavior comes about and inform the understanding of that behavior. The authors design the model this way to understand: when a target behavior can and cannot evolve; what gives it or prevents it from having adaptive value;

what are the consequences of its presence or absence, and, in particular, whether its presence confers a utilitarian advantage on the population having it.

According to the authors only the consequences for future utilities matter to the utilitarian judgment, therefore the utility function is the decision maker's best guess about how to maximize utility over the population. In this approach to utilitarian ethics, ethical problems and their solutions are matters of individual judgment. Each individual agent is confronted with choices and must choose an action amongst them at each decision point in its life, but they do not have direct access to anyone else's utility functions.

The following section describes the application of this model into different simulations which investigate how the fitness of different behaviors relate to its utilitarian ethical value.

3.2.2 Simulation

This subsection describes different simulations that use the model previously described and address different ethical issues, such as suicide, rape and abortion. The ethical issues addressed in these simulations are: the evolutionary status of altruistic suicide; the moral status of rape and its relation to dimorphic parental behavior from an act utilitarianism point of view; the moral status of abortion from an act utilitarianism point of view. The choice of such challenging and controversial topics demonstrates the possibilities of the implementation of ABMs in ethical research, reveals the potential value of simulation for understanding the value and morality of different behaviors [89], and opens new pathways for future research by inviting others to tackle such ethical issues better. The following are taken from [56, Chapter 5-6]. Tables listing the parameters used in each simulation can be found in the appendix 6.1.

Suicide as an Evolutionarily Stable Strategy This simulation is introduced in [56, p. 135-145]. Parameters used for the simulation are listed in the table 6.1. Suicide has adaptive value, that is, in conditions of high age and food deprivation and where suicide has a fitness benefit, it allows remaining agents better access to food, better reproductive potential, and better health. Suicides were conditioned upon at least low food density and high age, implying that removal of the agent from the simulation allowed other group members, who were more likely to be able to reproduce in the future, better access to food, and so suicide produced greater average health within the group, allowing remaining agents to better cope with the drought (in scenarios with seasonal drought). In these particular, circumstances suicide has adaptive value. Those circumstances were, of course, designed to provide a fitness benefit for suicide, with reliably repeated situations of environmental stress. Evolved altruism, of which suicide is only an extreme example, is widespread, both naturally and virtually.

Rape and Sexually Dimorphic Behavior This simulation is introduced in [56, p. 181-205]. Parameters used for the simulation are listed in the table 6.2. Utilities and health effects associated with the outcomes of the different actions are listed in the tables 6.3, 6.4, 6.5 and 6.6. The ethical investigations of rape show what is already known: rape is unethical. In the simulation, rape refers to any non-consensual act of mating. The parameter rape prevention probability (rpp) specifies the probability that a rape attempt will be repelled. When the (rpp) is high, both the level of rape and the level of its sexual dimorphism that evolve are low. It is assumed that victims, their families, friends and communities all are subjected to large negative utilities; and whatever happens to perpetrators, the ethical implications of their outcomes are dwarfed by the other consequences of their actions. Setting rape as an available action with large negative utilities for being a victim is never successful and always punished. If the action of raping is not available, the outcome is the most ethical

environment in all simulations. If rape becomes the fitter option for reproduction, it still results in a worst utilitarian outcome. Even in the case where rape is defined as an available action without negative utilities for being a victim, the outcome is worse than simulations without rape. The differential in health outcomes - larger cost in health for females, thus they must adopt less optimal strategies - for rape is plausibly a feature common to the general circumstances of rape, this supports the idea that rape cannot be turned into even an ethically neutral event by neutralizing its direct utilitarian costs. It is unethical at a more fundamental level.

Abortion This simulation is introduced in [56, p. 206-233]. Parameters used for the simulation are listed in the table 6.7. Utilities and health effects associated with the outcomes of the different actions are listed in the tables 6.8 and 6.9. Experiments looking into the utilitarian implications of abortion produced mixed results. In constant food simulations, abortion has no effect or a minor negative effect. In periodic drought scenarios, however, the effects of abortion range from a negative effect when the amount of investment needed is low (after-birth investment: $abi = 20$) to a positive effect when required investment is high ($abi = 150$ and $abi = 300$). In the latter cases, the utilitarian value of abortion corresponds to its adaptive value.

CHAPTER 4

Models of Ethical Issues

This chapter describes simulations of different issues related to ethics, thereby presenting the current state of the research in computational ethics by introducing several models related to ethical issues. These issues are tackled following different approaches, which in some cases overlap. They are grouped into different sections starting with agent-based models, followed by models related to game theory, and then to logic programming. These last two sections also contain a short introduction and motivation about their relation to ethics. The chapter is concluded with a section that presents other approaches that could not be categorized as neatly. Each subsection is structured in such a way that the simulation is introduced, its purpose described and the results presented. Followed by the specification of the model or models used. Each section is concluded with a short mention of the addressed ethical issue to summarize and motivate the model application to ethics.

4.1 Agent-Based Modeling and Ethics

This section introduces different simulations using ABMs related to ethics, which are not described as thoroughly as the representative example, rather explored enough to understand them and possibly compare them with one another. These subsections are structure such that there is an introduction of what is being simulated, followed by the specification of the model and concluded with a short mention of the ethical issue being addressed.

4.1.1 The Dynamics of the Evolution of Altruism

The model and simulation described here is introduced in [52]. This model focuses on the dynamics of the evolution of altruism and aims to simulate the evolutionary processes that led to the altruistic behavior. A variety of inheritable traits relating to altruistic action are present in the initial model population. The evolutionary fitness of a genetic trait is measured in relation to other traits, therefore the different altruistic traits are compared to a control group composed of non-altruists and unconditional altruists. Based on the changes in the distribution of the traits amongst the population over hundreds of generations, the sustainability of the different traits is benchmarked. The model also supports complex social networks, which allows for the development of relationships and group dynamics to be tracked and analyzed. The results show that:

- different conditions promote different types of altruism
- altruism leads to a considerably better survivability in harsh environments
- higher individual sacrifice does not improve the survivability
- cheaters destabilize altruistic systems significantly
- altruists lose local influence while gaining global influence

Models The altruistic behavior is represented by giving agents the opportunity to share some of their food supply with other agents that would otherwise starve. Giving food away hurts the agent's evolutionary fitness, since food is also an important requirement for reproduction. The model is separated into multiple submodels that model specific processes. Together, the BaseModel, AgingModel, ReproductionModel and EatingModel are the basis of the simulation by modeling the lifecycle of a population. Due to the object-oriented nature of the simulation, every model inherits properties and functions from its parent models. Different types of altruism are modeled by the children of the AltruismModel, which are:

- GreenbeardModel: altruism based on reciprocity. Agents will act altruistically towards others they believe to be greenbeards as well if given the chance, so if they belong to the same group;
- KinSelectionModel: relies on the genetic relatedness between agents to determine whether they would be willing to help each other in times of need
- ReputationModel: agents carrying the altruistic gene are willing to help those whose reputation is higher than the average reputation in the population. Acting altruistically also increases reputation. Also based on reciprocity
- GroupModel: agents are only willing to help those who are part of their group
- CultureModel: Extension of GroupModel. Each group has a culture value that determines the overall willingness to help each other. This changes over time depending on how the agents act.

The decision to act altruistically depends on the willingness to help, which is specified for each submodel depending on its representation of altruism, so it can be unconditional, based on reciprocity, reputation or group status.

Addressed Ethical Issue: How do different circumstances affect the evolution of altruism?

4.1.2 Investigating the Implications of Altruistic Behavior on Group Stability

The model and simulation described here is introduced in [83]. This model analyzes the impact of altruistic acts as costly practices on group stability while respecting the requirements of Darwinian evolution. It compares two groups, an altruistic one and a non-altruistic one, by including external living conditions that influence the fitness of the populations. The ABM is a NetLogo[1] model whose specification describes how altruism is represented in the model, how agents interact, how they are affected by external conditions, and how they change over time. The simulation is executed with different parameter setups where altruism and threats to the agents are respectively enabled and disabled. The results show

[1] NetLogo is a multi-agent programmable modeling environment for simulating natural and social phenomena. Reference: [103]

that altruistic groups survive under more severe and extensive external conditions than non-altruistic groups only if the benefit of an altruistic act far exceeds its cost. In milder circumstances, altruistic groups can have disadvantages regarding group stability compared to non-altruistic groups. [83]

Model The two groups are identical at the beginning of the simulation. The first group is the test group where agents can act altruistically. The second group is a copy of the first to ensure that both groups have the same starting conditions. Every tick corresponds to one generation of individuals. With every generational change, the fittest members of each group asexually procreate. Every tick, the simulation undergoes the following:

1. the initially set number of group members that is affected (*external-threat-scope*) loses the specific amount of fitness according to the initially set *external-threat-intensity*.
2. every agent of the first group whose fitness is high enough (*value* $\geq$ *altruism-fitness-threshold*) may act as an exemplar acting altruistically. The "learner" - the agent of the first group that receives help - is the member with the lowest value of fitness in the group. When the exemplar helps the learner, the learner's fitness is increased by the amount specified by the *altruism-fitness-gain*, the exemplar's fitness is decreased by the amount specified by the *altruism-fitness-cost*, and the learner's altruism is increased by 10.
3. the *procreation-percentage* fittest agents of each group are selected for procreation. Every selected agent gets *num-children* offspring. Amongst altruists, the altruism value is included into selection. Children inherit *group-index*, altruism value and fitness from their parent, though fitness is affected by mutation. Agents randomly gain or lose fitness according to the initially set *mutation-rate*. After procreation, the old generation dies.
4. count number of members and calculate average fitness of each group. For the altruistic group calculate the average altruism as well.

Addressed Ethical Issue: Impact of altruism as a costly practice in group stability.

4.1.3 Stability of Groups with Costly Beliefs and Practices

The model and simulation described here is introduced in [102]. *Costly signaling theory* proposes that animals may send honest signals about desirable personal characteristics and access to resources through costly biological displays, such as altruism, or other behaviors that would be hard to fake [57]. Henrich's cultural evolutionary model of costly displays shows that there can be a stable equilibrium for an entire population committed to costly displays, persisting alongside a no-cost stable equilibrium for the entire population [49]. This model is built in NetLogo and is a generalization of Henrich's results to a population peppered with subgroups committed to high-cost beliefs and practices, and aims to answer whether the same assumptions would yield similar equilibrium dynamics in a model that includes group identities in the simplest way possible. Agents use success-weighting calculations to determine whether to join or leave high-cost groups. According to the model, high-cost groups achieve long-term stability within a larger population under a wide range of circumstances. The most important emergent pathway to costly group stability is the simultaneous presence of high charisma and consistency of the group leader and high cost of the group. [102]

Model The base model (Henrich's model) is a cultural evolutionary model with replicator dynamics and discrete variables to capture belief and practice states. This model is an incremental extension where it is investigated whether the same assumptions would yield similar equilibrium dynamics in a model that includes group identities. This model requires agents that possess the same cognitive, communicative and interactive capabilities as the population members in Henrich's group-free model, but who can organize themselves into high-cost groups. That is, agents are designed with characteristics relevant to group dynamics and decisions about joining and leaving groups. The characteristics — charisma, consistency, sensitivity, tendency to affiliate with a high-cost group, and tendency to leave a high-cost group — are grounded in social theory. Most agents in the model can change dynamically, based on the variables describing their characteristics and tendencies. These variables are normally distributed across the population. The different types of agents can have 20 different types of encounters, each having a set probability and impact factor. In each encounter agents perform success-weighting calculations, which modify the agents' tendencies. The success-weighting calculations are formulas specific to each encounter type, they are essentially fitness calculations. Afterwards, group affiliation decisions are made based on thresholds on the characteristics and updated tendencies of the agents. Groups can die out or split based on specific thresholds.

Addressed Ethical Issue: Under which circumstances do groups of individuals with costly beliefs achieve group stability?

4.1.4 Agents with Values and/or Norms

The model and simulation described here is introduced in [61]. Agents with values and norms lead to simulation results that meet human needs for explanations. Here, agents with values and norms are modeled in a psychological experiment, and game theory game, on dividing money (pie): the ultimatum game, shortly UG, which demonstrates the reluctance to accept injustice. In the UG, two players negotiate over a fixed amount of money. The first player, the proposer, demands a portion of the money and offers the remainder to the other player, the responder, whom can choose to accept or reject this proposed split. If the responder chooses to *accept*, the proposed split is implemented. Otherwise, if the responder *rejects* the offer, both players get no money. The simulation outcome is then compared to empirical data on human behavior. Values serve as a static component in the agent behavior, whereas norms serve as a dynamic component. This two models are compared to one another and then to two others, the former being an extension of an already existing reinforcement learning model, the Learning Homo-Economicus model; the latter a combination of the first two, named Value-Norms model. The simulation is run in two different scenarios: Single-round scenario, where the reproduction of human behavior is investigated by letting the agents evolve and converge to stable behavior; Multi-round scenario, where the reproduction of the change in behavior humans display over multiple rounds of UG-play is investigated by different agent models. The agent model with values and norms produces aggregate behavior that falls within the 95% confidence interval wherein human behavior lies more often than any other tested agent models. [61]

Models Four models are compared with one another, in each model the agents behave differently:

- Learning homo economicus agent: only cares about maximizing their own welfare and can learn that forgoing short-time welfare might lead to a higher long-term welfare. Pie-portions are assigned utilities by the players, fir the proposer they represent the demand, for the responder the threshold. In the first round these initial utilities

are all equal to each other. There is a one-to-one relation to the utility of a pie-portion and the sum of the rewards that obtainable by the players.

- Value-based agent: the only relevant values in this UG are Wealth and fairness. The importance attributed to one or the other are perfectly negatively correlated. The higher wealth is valued, the higher the demands made and expected. The higher fairness is valued, the more equal the demands made and expected. The extent of satisfaction of these values is compared by a set of divide, product and difference functions.
- Normative agent: norms have four elements referred to as the 'ADIC'-elements: Attributes, to whom the statement applies; Deontic, permission, obligation or prohibition; Alm, action of the relevant agent; and Condition, scope of when the norm applies. Proposers expect that responders accept their demands, but reject everything higher than that. Responders expect that proposers demand the average of the lowest demand that is rejected and the highest demand that is accepted. If no norms exist, then players draw a random action from a uniform distribution.
- Value-based and normative agent: a combination of agents described in the previous two models. Some players always act according to the norm and others always act according to their values.

Addressed Ethical Issue: Does simulating agents with values and norms help to represent more closely human behavior (in the particular scenario of an UG)?

4.1.5 Relationship Between Culture, Values, and Norm Acceptance and Compliance

The model and simulation described here is introduced in [34]. The model comprises a population of agents characterized by a set of values and their ordering. Values are defined as ideals worth pursuing, which may be conflicting, and not valued equally by each individual. Each agent has an ordering over their values which, establishes their value profile. Norms are defined as standards within a society, which are aimed at achieving certain values. And culture is defined as the aggregation up to society level of the value profiles of the individuals within a given society. Values ordering is responsible for the agents' decision regarding their location and their interactions with other agents. The environment in which agents interact represents a public venue, like bars, pubs, cafes, etc. The modeled effect is smoke prohibition in cafes and the resulting agent behavior. The object of study of this simulation is related to the social perception of cigarette smoking in public. This target behavior is subject to the constraints coming from the general population of agents (social norms) or from the legal authority (legal norm). These two types of norms are defined as follows

- social norms: depends on the attitudes of the individual agents present in a specific venue, if the majority is in favor of smoking, the target behavior will be considered socially acceptable;
- legal norm: is introduced exogenously at run-time, half-way through the simulation, that is announced to every agent and can potentially conflict with the social norm in force at the moment of the introduction of the ban and/or with the individual attitude towards smoking.

Two models are presented, one where culture is modeled using norm type preferences; and a second one, where it is modeled in terms of values. The aim of the models is to show that culture makes a difference for policy effectiveness. The results of the first simulation, where

the values to a preference of complying to personal, social and legal norms are reduced, are that a relative small percentage of people not accepting smoking behavior, this rejection spreads and positively influences the uptake of the policy. In the second simulations, in which values are related to decisions taken over actions, seem to comply with the intuitions on how culture will influence the effect of the smoking ban.

Models In both models, agents have values and an ordering on these values that guides the decision regarding their location and interaction with other agents. They can interact in different public venues.

In the first model, each agent has one preferred norm type:

- lawful agents: law-abiding, whatever the law prescribes, they do
- social agents: whatever most of the agents in a certain shared context prefer, they do as well
- private agents: irrespective of law or context, they do what they themselves judge to be right

Legal norms range over entire society, social norms are relative to the agents present in the venue. Other than that, agents can either be in the venue or not and have a personal attitude towards smoking in venues. The decisions of the agents to enter/leave venues and their attitude towards smoking are coded in if-statements whose condition is based on thresholds determined by the preference of the majority of agents in the venue, private preference, type of norm followed. These preferences are randomly distributed. A complete order over the three norm types, contributes to a richer representation of different attitudes towards rules of conduct. The ratio in which each of the agent types is present in a society, reflects its culture with respect to rules of conduct.

The second model extends the previous one by including compliance. This is done by introducing a potential change in behavior, by linking attitudes to behaviors and showing how these behaviors are connected to agent values. Agents have discrete set of behaviors that define their possible choices. If they want to smoke, they decide whether want to do it in the social context they find themselves, based on their personal attitude towards smoking, the prevalent social norm in the venue, and whether a law banning smoking in public spaces has been enacted or not.

The simulation proceeds as follows: a population of agents, each time from a different cultural group, has the option to attend one of three possible venues. Their behavior inside the bar will determine the social norm related to smoking with a majority rule: if the number of smokers is higher than the number of non-smoking agents, then the venue will be said to have a social norm in favor of smoking. In the middle of the simulation a smoking ban prohibiting smoking inside public places is enacted. The introduction of a new legal norm is hypothesized to affect differently each cultural group. Each simulations cycle agents make three types of decision:

- select venue: agents choose to enter a bar or not, visit the same bar regularly or visit more bars based on randomly distributed thresholds or based on own attitude and venue's social norm
- smoke: the smoking behavior depends on the importance attributed to the others cultural values, the attitude towards smoking and the social norm prevalent in the bar, and, in the case of the enactment of a smoking ban, the presence of a legal norm that forbids smoking
- leave venue: depends on whether the agent's attitude regarding smoking is different from the social norm prevalent in the bar, or whether it is different from behavior

of the majority of the other customers. Once a smoking ban is enacted, the relative importance of the value of health will cause an agent to leave a smoker-friendly bar

The results show that different cultures react differently to a change in the legal norms.

Addressed Ethical Issue: How do culture, values, and norm acceptance affect individual behavior regarding unethical behavior such as smoking?

4.1.6 SimDrink: Simulation of a Night Consuming Alcohol

The model and simulation described here is introduced in [81] and [80]. SimDrink is a model build in NetLogo that simulates a population of 18-25 year old heavy alcohol drinkers on a night out in Melbourne and provides means for conducting policy experiments to inform policy decisions. The model consists of individuals and their friendship groups moving between private, public-commercial (e.g. nightclub) and public-niche (e.g. bar, pub) venues while tracking their alcohol consumption, spending and whether or not they experience consumption-related harms (i.e. drink too much), are involved in verbal violence, or have difficulty getting home. Individuals' behavior and decisions are setting dependent and allowed to vary as the night progresses. This variation is influenced by their and their friends' alcohol consumption, finances and harms experienced. This model is used to test and quantify the direct and indirect effects of policies such as 24 hour public transport, public venue lockouts, changes to responsible service of alcohol enforcement, public venue closing times and drink prices. [81] Results:

- a two-hour extension of public transport is likely to be more effective in reducing verbal aggression and consumption-related harms than venue lockouts
- Modeling a further extension of public transport to 24 hours has minimal additional benefits and the potential to displace incidents of verbal aggression amongst outer urban area residents from private to public venues.
- When implemented in conjunction with any extension of public transport, 3am lockouts were as effective as 1am lockouts in reducing verbal aggression.

The model's relevance in ethics research is related to the fact that it is oriented toward inhibiting morally problematic behaviors such as alcohol abuse. The model uses different probabilities based on qualitative studies to model the agents' decision making, these probabilities are different for each specific circumstance. For example $p\_PTrush\_OU\_plan\_\$$ describes the probability that an individual will choose to catch the last train home if they have less than $50 left, had only planned to stay out for up to one hour longer and live in an outer urban area. [80]

Model The environment consist of the inner city area, where everyone goes to party, and the outer urban area, which consists only of private venues. Agents move in the environment, consume alcohol without exceeding their personal drinking limit and money, belong to friendship groups, and can experience different harms: verbal harms, drinking too much, difficulty getting home. The venues they visit have closing times if they are private. Each time step, public venues can eject intoxicated patrons or close, agents can move between venues, agents can consume drinks, agents determine harms experienced, agents can consider going home, and get home by taxi, and friendship groups are activated. The decisions taken by the agents are mostly described in if-statements - each possible action is described - which also contain different probability distribution for each specific circumstance. These distributions are based on qualitative studies about young people's drinking

events. The simulation parameters are taken from publicly available information for Melbourne about public transport and alcohol consumption, from studies about young adults' alcohol consumption, from available literature, or they are plausible estimates made by the authors based on their experience in social research on alcohol and other drug use in the night-time economy. These parameters are tested in a sensitivity analysis and as part of a Latin Hypercube uncertainty analysis.

Addressed Ethical Issue: How does policy making affect the inhibition of morally problematic behaviors such as alcohol abuse?

4.1.7 Generative Explanatory Model of Offending Behavior: a Simulation of Residential Burglary

The model and simulation described here is introduced in [23]. This model simulates residential burglary in a world inhabited by potential targets and offenders who behave according to the theoretical propositions of environmental criminology. The simulation examines how different mechanisms impact patterns of offending. This models tests whether the proposed micromechanisms of the routine activity approach, rational choice perspective, and crime pattern theory are generatively sufficient to produce three macroscopic regularities of crime hot spots, repeat victimization, and journey to crime curve. The following hypothesis are investigated with these mechanisms enabled:

- Crime will become more spatially concentrated
- Greater levels of repeat victimization will be observed
- The journey to crime curve will become more positively skewed

The model environment is a two-dimensional grid containing navigational nodes, potential targets and potential offenders. Offenders move to encounter targets, follow a decision making process and become aware of the spaces they visit. The outputs of these simulations then are compared with several findings derived from empirical studies of residential burglary, including the spatial concentration of crime, repeat victimization, and the journey to crime curve. The model suggests that with respect to those crimes that occur against static targets, of the three mechanisms examined, it is the spatial and temporal constraints of offender activities, and their subsequent knowledge acquisition, which have the greatest impacts on patterns of spatial clustering, repeat victimization, and the journey to crime curve. [23]

Model The model environment is a two-dimensional grid containing navigational nodes, potential targets and potential offenders. Offenders have a routine activity space, which identifies their commonly visited locations throughout the environment. When encountering a possible victim, they go decide whether to act or not based on the metric of attractiveness of the potential target, which is a combination of its associated risk, reward, and effort represented in the interval $[0, 1]$, where 1 is a target with the greatest rewards, the smallest risk/effort, and 0 is a target with few rewards, considerable risk/effort. They also have an awareness space represented as a spatially referenced two-dimensional matrix of awareness scores between 0 and 1, mapping directly to the environmental lattice. Offender awareness changes. As offenders move across environment, their awareness of visited locations increases. The learning rate is selected such that offender awareness of a given environmental lattice approaches 1 after it has been visited 50 times. The likelihood of a crime occurring is defined as product of the probabilities that a suitably motivated offender finds a sufficiently attractive target of which they are sufficiently aware.

Addressed Ethical Issue: What has the biggest impact on patterns of spatial clustering of crime, repeated victimization, and the journey to crime curve?

4.1.8 Generative Model of the Mutual Escalation of Anxiety Between Religious Groups

The model and simulation described here is introduced in [87]. A generative model of the emergence and escalation of xenophobic anxiety in which individuals from two different religious groups encounter various hazards within an artificial society. The model generates mutually escalating xenophobic anxiety between two religious groups under theoretically sound conditions that are consistent with Terror Management Theory (TMT), Social Identity Theory (SIT), and Identity Fusion Theory (IFT). Mutually escalating xenophobic anxiety occurs when the average anxiety level of agents in both groups increases overtime. The model uses decision trees for the interactions and decisions made by entities within the model at each time step. The trace validation techniques used show that the most common conditions under which longer periods of mutually escalating xenophobic anxiety occur are those in which the difference in the size of the groups is not too large and the agents experience social and contagion hazards at a level of intensity that meets or exceeds their thresholds for those hazards. Under these conditions agents will encounter out-group members more regularly, and perceive them as threats, generating mutually escalating xenophobic anxiety. The results show that the most common conditions under which longer periods of mutually escalating xenophobic anxiety occur are those in which the difference in the size of the groups is not too large and the agents experience social and contagion hazards at a level of intensity that meets or exceeds their thresholds for those hazards.[87]

Model

- TMT: two of the most common consequences of death awareness are increased acceptance of the existence of hidden intentional force, especially supernatural agents - defined as anthropomorphic promiscuity by the authors - and increased resistance to engaging other cultures - defined as sociographic prudery by the authors.
- SIT: social identity are aspects of an individual's self-image that are shaped by their sense of belonging to a particular social category. This theory hypothesizes that pressures to evaluate one's own group positively through in-group/out-group comparisons leads social groups to attempt to differentiate themselves from each other [96] The interaction between groups can be determined by value laden social differentiations that ratchet up tension between the groups, which can then lead to conflict and violence [95] This theory focuses on the double role of group membership and social categorization in shaping group cohesion and contributing to intergroup conflict.
- IFT: Extreme identity fusion involves the blurring of personal and social identities such that the group comes to be regarded as functionally equivalent to the self. Identity fusion is a distinctive construct that refers to how individual identity interacts with group identity in a synergistic and reinforcing dynamic [44] Less fused people may have strong beliefs about what sacrifices "ought" to be made for their group, but are less likely to act on those beliefs compared to highly fused people, who are more willing to kill, or even die, for the group [93] [100].

Agents are subjected to different types of hazards that increase their stress and heighten their mortality salience. These encounters can provoke members of the groups to seek explanations and help from supernatural agents, and to turn to fellow group members for comfort and protection, thereby increasing their desire to engage in shared rituals (as

predicted by TMT). As these ritual engagements intensify, some agents become more fused to their in-groups, which increases their propensity towards feeling anxious about out-group members (as predicted by SIT and IFT). Variables:

- Independent variable: simulated heterogeneous agents distributed in two groups in an artificial society
- Intervening Variables: group size and threat variables are altered and agents interact based on TMT, SIT and IFT
- Dependent Variables: Mutually escalating xenophobic anxiety between religious groups

Addressed Ethical Issue: Under which conditions can escalating xenophobic anxiety occur?

4.1.9 Terror Management Theory

The model and simulation described here is introduced in [86]. Following the description of two models designed to simulate the dynamic systems and behavioral patterns identified and clarified by research on terror management theory (TMT). TMT asserts that the human reaction to the anxiety evoked by awareness of death involves the construction and maintenance of cultural worldviews that help to bolster self-esteem, psychological equanimity, and a sense of meaning. The first model is a System-Dynamics Model (SDM), which attempt to formalize the causal architecture of a complex non-linear system using stocks - storage of some units in the model -, flows - movement of units between stocks -, time delays, and the interactions of variables that may form positive or negative feedback loops. The second is an ABM that extends the first. These are modeled as causal architectures informed by empirical research on the effects of mortality salience on "religiosity" (and vice versa). They are also informed by research on the way in which perception of personal and environmental hazards activate evolved cognitive and coalitional precautionary systems that can intensify anxiety-alleviating behaviors such as imaginative engagement with supernatural agents postulated within a religious coalition, shortly research on how perceived hazards augment religious belief. The aim of the simulation is to model the effect of mortality salience on religiosity and vice versa. Each of the simulations is focused on the relevant inputs and outputs that shape the ways in which, more or less religious, agents adapt to waves of threatening, anxiety-producing events, defined by the authors as "natural adaptation to hazard undulation models", shortly NAHUM. The NAHUM-SDM experiment explores ways in which an individual's personality traits can shape reactions to life-threatening environmental stimuli and alter their religiosity over time. The NAHUM-ABM experiments explore the role of mortality salience in the social interaction of heterogeneous religious individuals who intensify their ritual engagement in response to environmental hazards. [86]

Model The goal of the model is to abstract some of the most religiously salient causal mechanisms studied in the literature in order to model them within computer simulations and provide another way of validating them experimentally. The authors define the following:

- anthropomorphic promiscuity: hyperactive propensity towards detecting gods as hidden agents
- sociographic prudery: hyperactive propensity towards protecting in-group norms
- Religiosity: socially shared cognitive and ritual engagement with axiological relevant supernatural agents postulated within one's in-group

In the first model, NAHUM-SDM, stocks indicate levels of religiosity, flows indicate the movement of religiosity between stocks and variables are either fixed or dynamic and have different effects on either stocks or flows. These variables can be personal, related to simulated individual, or environmental, related to four types of hazard - social, contagion, natural, and predation. The hazards values can change, encounters with social and contagion hazards increase sociographic prudery (SP), while natural and predation hazards increase anthropomorphic promiscuity (AP). Environmental variables have two aspects, occurrence rate and intensity, that increase religiosity outputs, which decay over time at a specific rate: religiosity decay, that is the rate at which an individual's heightened levels of AP and SP decay over time, the lower religiosity decay rate, the slower religiosity decays after a threat; habituation rate, that is the rate at which an individual's reliance on religious ritual to mitigate threat-induced stress declines., the lower habituation rate, the slower belief in the efficacy of ritual interventions to manage terror declines, the longer it takes to become habituated to the threatening event. The experiment is designed such that the conditions under which the system can detect the emergence of these four trends over time is explored:

- maintenance of AP and SP
- steady increase of AP and SP
- steady decrease of AP and SP
- cycling between low and high levels of of AP and SP

This model can reproduce all four targeted behavioral trends, offers a formalized computational model that can be edited and expanded, provides a new experimental tool for studying the dynamics of the personal characteristics and relations of religious individuals as they react to various environmental threats of different intensities and at different rates of occurrence.

The second model, NAHUM-ABM, has rules, which agents follow in each run, specify how agents react personally to perceived threats as well as how they interact with other agents. The model is initialized with 100 agents assigned to one of two groups. Agents act differently with other in-group and out-group agents. They encounter various hazards: social or contagion threats associated with out-group members, and predation and natural threats associated with the environment. They differ in their capacity to tolerate threats, which affects the extent to which they react to such hazards. Agents whose stress is exacerbated by mortality salience will tend to cluster with other agents of the same type and intensify their performance of rituals, which helps to alleviate the stress caused by the threat. Ritual in-groups never contain agents of more than one type and are composed of agents with similar levels of religiosity. As agents continue their ritual engagement, their level of religiosity will tend to become more like the (averaged) religiosity of those agents with whom they have just ritually interacted. In any given run, therefore, an agent's levels of AP and SP (i.e., religiosity) change over time, which in turn affects the other agent's perception of them as viable ritual co-participants. Most of the variables are the same as in NAHUM-SDM, with the addition of group-level, which is the ratio of the number of members in the two groups and allows to simulate the relationship between majority and minority groups within a population; ritual cluster size, which can be tracked as agents gather together with in-group members with a similar religiosity level to perform rituals in the face of threats. At an individual level, agents vary in their tolerance levels for, and their susceptibility to becoming stressed by, each of the hazards. The simulation's goal is to replicate findings in the TMT literature, which helps to validate the causal architecture. Investigate way in which the level of prior religiosity (with high prior religiosity, meaning a high sum of AP and SP, being a proxy for fundamentalism) affects the extent to which an individual's religiosity is maintained or increased during a simulation run. The simulation

is designed to explore the social dynamics within and across religious groups as a population encounters various environmental hazards, and focuses on sizes of the ritual groups that formed within the entire population during a simulation run, altering personal and contextual conditions in order to tease out relevant causal dynamics. The results demonstrate that individuals with high initial religiosity rely upon their religiosity to alleviate stress to a greater extent than individuals with low initial religiosity

Addressed Ethical Issue: What are the conditions that shape the ways agents adapt to waves of threatening, anxiety-producing events?

4.1.10 Prediction of Changes in the Existential Security and the Religiosity of a Group

The model and simulation described here is introduced in [45]. It employs existing data sets and ABM to forecast changes in religiosity and existential security amongst a collective of individuals over time. The model includes agents in social networks interacting with one another based on their education level, religious practices, and existential security within their natural and social environments. The inclusion of social networks with educational homophily - principle that a contact between similar individuals occurs at a higher rate than amongst dissimilar individuals [60] - improves forecast accuracy, which alters the way in which religiosity and existential security change in the model. These dynamics grow societies where two individuals with the same initial religious practices evolve differently based on the educational backgrounds of the individuals with which they surround themselves. [45]

Model The goal of the simulation is to use ABM to predict changes in the existential security and the religiosity of a collective of individuals, for a given time period and country. Where existential security is defined as the extent of economic, socioeconomic and human development provided by a country [66]. Each agent has an education level, an existential insecurity level, and four variables that reflect their religiosity, namely, religious formation, religious practice, supernatural beliefs, and belief in God. Each agent is also connected to a subset of the other agents in the model through a social network, and to the existential security level of the environment.

The existential security level of the environment reflects the percentage of the agents that feel the level of economic, socioeconomic and human development support provided to them is sufficient. An agent determines if they feel existentially secure by checking if their value for existential insecurity is below the existential security value of the environment. The existential security of the environment is parameterized by the Human Development Index - the Human Development Report (HDR) is an annual multifaceted analysis of well-being focused on key dimensions of human development including a long life, a healthy life, and a decent standard of living, and the HDI is the summary measure used in the HDR for a country's achievement across these dimensions [4].

The social network is generated using an algorithm for the generation of social network graphs, described in [32], which uses the Education Homophily Parameter - degree to which the educational level of the agent is correlated to the emotional closeness [37]. . The social network also influences the religious practice of the agent, which itself influences belief in god and supernatural beliefs, the latter being influenced by religious formation as well. These four, plus education level, are initialized by sampling respondents of the international social survey program, ISSP [33]. Structural equation modeling (SEM) is then used to organize the relationships amongst the four religiosity factors.

The model was evaluated against alternative modeling approaches, namely, the baseline approach, based entirely on historical data; and a statistical approach, which uses linear regression modeling.

Addressed Ethical Issue: Under which conditions does religiosity and existential security of individuals with same initial religious practices change over time?

4.1.11 Ethnonationalist Radicalization Between Political Actors and Their Constituencies

The model and simulation described here is introduced in [65]. The simulation models ethnonationalist radicalization between political actors and their constituencies based upon evidence from former Yugoslavia. The central mechanism is the recursive feedback between political and cultural dynamics, focusing on processes prior to the outbreak of actual violence.

As stated by based modeling strategy, the model's description is accompanied by a motivation of the model assumptions from the empirical evidence [40] [105]. The evidence is as descriptive as possible in order to avoid theoretical preconceptions. The model's target is the processes of nationalist radicalization prior to actual violence. The emergence of the militia only serves as an indicator for the radicalization of political attitudes; the complexity of their operations is not represented. This focus suggests specifying the research questions as, first, how a political agenda resonates with its audience and, second, to determine the tipping point at which the micro dynamics of nationalist escalation processes become self-perpetuating.

Taking the recursive influence between political actors and constituencies into account enables the questioning of the diversity-breeds-conflict theory: political radicalization and counterradicalization are more likely in ethnically homogeneous societies with a common historical legacy such as Croatia and Serbia. Initially, ethnically mixed societies such as Bosnia–Herzegovina are less likely to fall victim to political radicalization.

The results offer theoretical insights by revealing mechanisms that lead to escalation. Multiethnic regions are more capable of withstanding political pressures, but vulnerable to imported violence, driven by the local population. [65]

Model The model's target is the processes of nationalist radicalization prior to actual violence. The emergence of the militia only serves as an indicator for the radicalization of political attitudes; the complexity of their operations is not represented. This focus suggests specifying the research questions as, first, how a political agenda resonates with its audience and, second, to determine the tipping point at which the micro dynamics of nationalist escalation processes become self-perpetuating. There a two types of agents: citizens, who have different civil values and national identity, which can change over time; and politicians, who want to promote their own career, act by giving civil, radical nationalist, or moderate nationalist speeches. The likelihood of changing the agenda in the next round is determined by a strategic evaluation of feedback from the citizens about: political climate, credibility and exclusiveness.

A state of complete ignorance is assumed for the cognitive components of the agents. During the simulation, the value orientation of the citizens changes due to the mobilizing influence of politicians' speeches. Constituencies can choose to support different speeches supplied by different politicians. To maximize their public support, politicians adapt their speeches to the value orientations of the constituencies. The steps followed are:

1. political mobilization: politicians make speeches to galvanize support. Individuals can strongly support a speech. Or, individuals discuss the speeches within networks

of neighbors, changing the initial evaluation. After this discussion, people can decide to participate together at a demonstration.

2. political conflicts: when a nationalist politician gains support outside the territory of the home republic, political conflict intensifies. The reaction can be that a modest alarm is activated, which increases the likelihood that the politician will moderate the speech. If not, a stage of wholesale political conflict is reached, which may give way to military action.
3. conditions for violence: opportunities, motivation, complicity
4. anomic (socially disorganized) system state: if these conditions are fulfilled, the militia carries out ethnic cleansing within its local neighborhood, including murder and displacement. Survivors flee in the direction of their friendship network. Refugees are modeled as strongly radicalized nationalists: If they find collaborators, they accompany a militia.

Addressed Ethical Issue: Under which scenarios can the risk of escalation of ethnonationalist radicalization occur?

4.1.12 The Virtue of Temperance

The model and simulation described here is introduced in [54]. The model is an extension of an ABM from NetLogo, called Sugarscape, with the addition of the cognitive architecture PECS to represent the decision making process of the agent in simulating the virtue of temperance. The virtue of temperance is self-control over natural physical desires (e.g. moderation in matters of food). PECS is a component-oriented agent architecture which enables integrative modeling of physical conditions, emotional state, cognitive capabilities, and social status. Here it is used to determine whether the agent will pursue or ignore the food. Respecting cognitive rules and not breaking social rules allows the agent to ”learn” the virtue of temperance. [54]

Model The model environment is a two-dimensional grid containing agents and food. The agents have a starting amount of sugar, metabolize sugar at every turn and if no sugar is left the agent dies. They can also move around grid to survive, have variable range of vision, their choice of next grid cell to move is restricted to their vision range, patch availability and amount of sugar. They aim to be healthy. The model is defined such that eating three sugars is unhealthy and has the social rule ”it is wrong to eat more than 1 unit of sugar at a time”. The food in the environment is represented by sugar patches, which can have different amounts of sugar, and after the contained sugar is consumed, it can grow back according to predetermined regrowth rates. Agents go through a specific decision process and change their decisions based on punishments from social and physical rules, and rewards from cognitive and emotional rules.

Addressed Ethical Issue: Does the virtue of temperance improve the conditions of a group?

4.2 Game Theory and Ethics

This section describes how game theory is used in computational ethics and introduces models of ethical questions which use game theory approaches.

Game theory is the study of interdependent choice and action. It includes the study of strategic decision making, the analysis of how the choices and decisions of a rational agent depend on (or should be influenced by) the choices of other agents, as well as the study of group dynamics, the analysis of how the distribution of strategies in a population evolves in various contexts and how these distributions impact the outcomes of individual interactions [47]. It can be used for analyzing and predicting how humans behave in strategic situations, by using three distinct concepts to make precise predictions of how people will interact strategically: strategic thinking, all players form beliefs based on an analysis of what other players may do; best-reply, choice of best reply given specific beliefs; and mutual consistency (equilibrium), adjust the best responses and beliefs until they reach an equilibrium. [30] When applied to ethics research, game theory can assist to establish the functions of morality, i.e. to assist in preventing failures of rationality (functionalist approach), formalize social contract theory (contractarianism approach), or to recover and establish the origin of moral norms (evolutionary approach). It provides a suitable framework that explains the emergence of norms used to guide agent behavior during the ethical decision making process [90].

The application of game theory to ethical decision making can be classified into the following [90]:

1. functionalist approach: use game theory to establish the functions of morality
2. contractarianism approach: use game theory to formalize social contract theory
3. evolutionary approach: use evolutionary game theory to recover and establish the origin of moral norms

4.2.1 Manipulation Based on Machiavellianism

The model and simulation described here is introduced in [30]. The model is specified with a game theory approach for modeling manipulation behavior based on Machiavellianism, which is a social conduct strategy supposing that the world can be manipulated by applying (Machiavelli's) tactics with the purpose of achieving personal gains according or not to a conventional moral. The model is build using a combination of the deontological and utilitarian moral rules. The players' acquisition of moral or immoral behavior is contributed by a reinforcement learning algorithm's principle of error-driven adjustment of cost/reward predictions, based on an actor-critic approach responsible for evaluating the new state of the system. The algorithm determines if the cost/rewards are better or worse than expected, supported by the Machiavellian game theory solution. The model simulates manipulation where the manipulator players can anticipate the predicted response of the manipulated players. Using a Machiavellian game theory approach, they establish a system that analyzes and predicts how players behave in strategic situations combining: strategic thinking, best-reply, and mutual consistency. To model Machiavellian immorality, the authors, use utilitarian and deontological moral theories and use reinforcement learning as mechanism by which utilitarian moral value may guide the Machiavellian behavior, and the deontological moral to obtain a value for certain moral principles or acts. The validity of this method is demonstrated theoretically and with a simulated numerical example. [30]

Model The following concepts are defined and used in the model by the authors:

- Views: The belief that the world can be manipulated. The world consists of manipulators and manipulated
- Tactics: The use of a manipulation strategies needed to achieve specific power situations (goals)

- Immorality: The disposition to not become attached to a conventional moral
- Machiavellian social conduct: manipulating others for personal gain, even against the other's self-interest [27]
- Machiavellian intelligence: capacity of an individual to be in a successful engagement with social groups [27]

In order to manipulate, here acquire power, survive or sustain a particular position, Machiavellian agents make use of Machiavellian intelligence applying different selfish manipulation strategies, which include looking for control the changes taking place in the environment.

The learning agent is split into the actor and the critic. The actor coincides with a Stackelberg/Nash game that implements the concepts of Views and Tactics of the Machiavellianism. The actor is responsible for computing a control strategy, for each player, given the current state. After a number of strategy evaluation steps by the critic, the actor is updated by using information from the critic. The action selection follows the strategy solution of the Machiavellian game:

1. given a fixed current state, each player chooses randomly an action from the vector of the strategy solution
2. then, players employ the transition matrix, of the probabilities associated with the transition from state to state under a specific action, to choose randomly the consecutive state from the vector of the probabilities associated with the transition from state to state under a specific action,
3. as soon as current state, actions and consecutive state are selected they are sent to the Critic.

After a number of strategy evaluation steps by the critic, the actor is updated by using information from the critic.

The critic conceptualizes the Machiavellian immorality represented by a value function process. The critic is responsible for evaluating the quality of the current strategy by adapting the value function estimate, and compute an approximation of the projection of the vector (of the probabilities associated with the transition from state to state under a specific action) and the average cost function. A Machiavellian player has the disposition to not become attached to a conventional moral as it has a combination of the deontological and utilitarian moral rules, as well as, moral heuristics, for representing the concept of immorality decision-making. Here deontological moral states that certain things are morally valuable in themselves - action is done to protect such moral values: consequences and outcomes of actions are secondary - whereas utilitarian moral establishes states where the moral quality of actions is determined by their consequences - maximize the utility of all individuals in the society.

The Machiavellian game dynamics are as follows, the manipulator players consider the best-reply of the manipulated players, and then select the strategy that optimizes their utility, anticipating the response of the manipulated players. Subsequently, the manipulated players observe the strategy played by the manipulators players and select the best-reply strategy. The manipulators and manipulated players are themselves in a (non-cooperative) Nash game. Formally, the Stackelberg model is solved to find the subgame perfect Nash equilibrium, i.e. the strategy that serves best each player, given the strategies of the other player and that entails every player playing in a Nash equilibrium in every subgame. The equilibrium point of the game represents the strategies needed to achieve specific power situations. In the model, the manipulators have commitment power presenting a significant advantage over the manipulated players. A reinforcement learning approach is employed for representing immorality. This provides a computational mechanism, in which,

its principle of error-driven adjustment of cost/reward predictions contributes to the players' acquisition of moral/immoral behavior. The reinforcement learning algorithm is based on an actor-critic approach responsible for evaluating the new state of the system and determine if the cost/rewards are better or worse than expected, supported by the Machiavellian game theory solution. The algorithm is viewed as a stochastic game algorithm on the parameter space of the actor. The game is solved employing the extraproximal method. The functional of the game is viewed as a regularized Lagrange function whose solution is given by a stochastic gradient algorithm. When the player's performance is compared to that of a player which acts optimally from the beginning, the difference in performance gives rise to the notion of regret. Then, the critic produces a reinforcement feedback for the actor by observing the consequences of the selected action. The critic takes a decision considering a temporal difference error, shortly TD-error, which determines if the cost/rewards are better or worse than expected with the preceding action. The TD-error corresponds to the mean squared error of an estimator which in this case measures the difference between the estimator and what is estimated - the difference occurs because of randomness. The temporal difference error in the reinforcement learning process is employed to evaluate the preceding action: if the error is positive the tendency to select this action should be strengthened, otherwise reduced. When the actor-critic learning rule ensures convergence, the value-minimizing/maximizing action at each state is taken.

Addressed Ethical Issue: What is the moral status of manipulation?

4.2.2 Evolving Agents with Moral Sentiments in an IPD Exercise

The model and simulation described here is described in [20], [19] and [18]. It simulates a society of agents where some of them have "moral sentiments" towards the agents that belong to the same social group. The Iterated Prisoner's Dilemma, shortly IPD, is used as a metaphor for the social interactions. [18] The agents in the IPD behave rationally, thus named egoistic agents, or have moral sentiments towards those from the same social group, named altruistic agents. Behaving rationally is not the best attitude for a good performance in the long run, both individually and for the group. The results show that:

- the more egoist agents, the worse the general performance of the society as a whole; that is, the fewer total points from all agents in all groups are accumulated in the same period;
- the more egoists in a group, the faster the group collects points initially, but the worse its eventual performance in the long run; the fewer egoists in a group, the better the group's performance;
- the smaller the percentage of egoists in a mixed group, the better the performance of each individual egoist agent

[20] [19]

Models At each play, a pair of agents is chosen randomly amongst agents of all groups. Agents earn points by playing, and, depending on group, have different values for temptation to defect, reward for mutual cooperation, punishment for mutual defection, and sucker's payoff. Altruistic agents have a wealth state, which is the average number of points through the completed simulation steps or through a determined period of time. Thus, these agents are classified based on wealth state into wealthy, medium and straitened status, according to specific thresholds. The two agent groups have the following strategies: egoistic agents always defect; whereas altruistic agents, use the TFT strategy - equivalent retaliation [14] - when playing against agents from other groups, and when

playing against agents from same social group, they use the Moral Sentiments strategy (MS), that is, an altruist cooperates with those of the same group unless it is in a straitened state and the opponent is wealthy. The simulation is set up such that it has societies of agents of 3 or 15 groups with 4, 20, 40, 80, 100 agents. In some simulations all agents play at a certain time, in others only 60% or 80% of agents play; The 3 groups are composed of only altruists, 50% egoists and 50% altruists, and only egoists.

The evolutionary approach of the model entails a population of agents reproduced according to their fitness, designed by a genetic algorithm, where recombination of strings are not allowed, resulting in asexual reproduction. Shortly, it is defined as k fittest individuals generate a number of offsprings, here copies of themselves, proportionally to their fitness.

The score of each agent is computed over the last 10 time steps. If the agents collects more than the threshold points it is considered wealthy, otherwise straitened. Agent characteristics are coded in binary strings: type, egoist/altruist; wealth state; group it belongs to. The reproduction frequency is set at beginning of simulation, and it defines how agents play, accumulate points, and die, which happens immediately after reproducing. Accumulated points determine the probability that they will reproduce. The number of agents in each group can increase/decrease and groups can disappear.

Addressed Ethical Issue: Do moral sentiments affect individuals' decision making?

4.3 Logic Programming and Ethics

This section describes how logic programming is used and how it relates to computational ethics, and introduces models of ethical questions which use logic programming approaches.

Logic programming in computational ethics is used as a vehicle for the computational study and teaching of morality, in its modeling of the dynamics of knowledge and cognition of agents, by studying norms and moral emergence in populations of agents. [78] Logic-based approaches have a great potential to model moral machines, in particular via non-monotonic logics. Ethical theories and dilemmas are often represented in a declarative form by ethicists, who also used formal and informal logic to reason about them. In addition to that, ethical rules are default rules, so they tolerate exceptions. Combining logic-based representation and logic-based learning for modeling ethical agents provides many advantages: increases the reasoning capability of agents, therefore influences their decision making process; promotes the adoption of hybrid strategies that allow both top-down design and bottom-up learning via context-sensitive adaptation of models of ethical behavior; allows the generation of rules with valuable expressive and explanatory power, thus equipping agents with the capacity to make ethical decisions, and to explain the reasons behind these decisions [39].

4.3.1 Agents that Judge One's Own and Others' Behaviors

The model and simulation described here is introduced in [31]. This model aims to producing ethical behaviors from a multi-agent perspective. This is a model of ethical judgment an agent can use in order to judge the ethical dimension of both its own behavior and the other agents' behaviors. The model is based on a rationalist and explicit approach that distinguishes theory of good and theory of right, and is a proof-of-concept model implemented in Answer Set Programming. It illustrates an ethical agent in a multi-agent system where agents have beliefs, about richness, gender, marital status and nobility; desires; and

their own judgment process. They are able to give, court, tax and steal from others or simply wait. [31]

Model According to the authors, ethics consists in conciliating desires, morals and abilities. To take these dimensions into account, the generic ethical judgment process, shortly EJP, uses evaluation, moral and ethical knowledge. It is structured along Awareness, Evaluation, Goodness and Rightness processes. EJP is considered in the context of a BDI model, using also mental states such as beliefs and desires. For simplicity reasons, ethical judgment reasoning is only considered on short-term view by considering behaviors as actions. This model is only based on mental states and is not dependent on a specific architecture. In more detail, an ethical judgment process, EJP, is defined as a composition of an Awareness Process, an Evaluation Process, a Goodness Process, a Rightness Process, an Ontology of moral values and moral valuations. It produces an assessment of actions from the current state of the world with respect to moral and ethical considerations. The model is a global scheme, composed of abstract functions, states and knowledge bases. These functions can be implemented in various ways. For instance, moral valuations from the ontology may be discrete such as $\{good, evil\}$ or continuous such as a degree of goodness.

- Awareness Process: generates the set of beliefs that describes the current situation from the world, and the set of desires that describes the goals of the agent
- Evaluation Process: produces desirable actions and executable actions from the set of beliefs and desires
- Goodness Process: identifies moral actions given the agent's beliefs and desires, the agent's actions and a representation of the agent's moral values and rules
- Rightness Process: produces rightful actions given a representation of the agent's ethics

Agent uses this process to judge its own behavior and that of others. The categories of ethical judgments based on the amount of information the agents has about the other agent being judged are:

- Blind ethical judgment: the judgment of the judged agent is realized without any information about this agent, except a behavior,
- Partially informed ethical judgment: the judgment of the judged agent is realized with some information about this agent,
- Fully informed ethical judgment: the judgment of the judged agent is realized with a complete knowledge of the states and knowledge used within the judged agent's judgment process.

Addressed Ethical Issue: How to judge self and others ethically?

4.3.2 GenEth: A General Ethical Dilemma Analyzer

The dilemma analyzer here is introduced in [8]. GenEth is a general ethical dilemma analyzer that, through a dialog with ethicists, uses inductive logic programming to codify ethical principles in any given domain and provide assistance in discovering ethical principles. More formally, a definition of a predicate p is discovered such that $p(a_1, a_2)$ returns true if action a_1 is ethically preferable to action a_2. The principles discovered are mostly general specializations, covering more cases than those used in their specialization and, therefore, can be used to make and justify provisional determinations about untested cases and are more general.

GenEth is committed to a knowledge representation scheme, based on the concepts of ethically relevant features, which have corresponding degrees of presence or absence, from which duties to minimize or maximize these features are inferred. These minimization/maximization is done with corresponding degrees of satisfaction or violation of those duties. The system has no a priori knowledge regarding what particular features, degrees, and duties in a given domain might be. It determines features, degrees, and duties in conjunction with its trainer as it is presented with example cases. The advantages of this approach are that it minimizes bias, and the principle in question can be tailored to the domain with which one is concerned. Different sets of ethically relevant features and duties can be discovered, through consideration of examples of dilemmas in the different domains investigated. Different features and duties can be added or removed if it becomes clear that they are needed or redundant. [8]

Dilemma Analyzer Description The ethical action preference is dependent on ethically relevant features that actions involve - harm, benefit, respect for autonomy, etc. A feature is represented as an integer that specifies the positive value, that is, degree of presence in a given action; or negative value, the degree of absence in a given action. For each ethically relevant feature, there is an agent duty to minimize that feature, for example harm; or maximize it, for example respect for autonomy. A duty is represented as an integer that specifies a positive value, that is, degree of satisfaction in a given action; or negative value, the degree of violation in a given action. An action is represented as a tuple of integers, representing the degree to which it satisfies or violates a given duty. A case relates two actions and is represented as a tuple of the differentials of the corresponding duty satisfaction/violation degrees of the actions being related to positive case - the duty satisfaction/violation degrees of the less ethically preferable action are subtracted from the corresponding values in the more ethically preferable action, producing a tuple of values representing how much more or less the ethically preferable action satisfies or violates each duty than the less ethically preferable action; or negative case - the subtrahend and minuend are exchanged. A principle of ethical action preference is defined as an irreflexive disjunctive normal form predicate p in terms of lower bounds for duty differentials of a case. Δd_i is the differential of the corresponding satisfaction/violation degrees of duty i in actions a_1 and a_2. $v_(i,j)$ denotes the lower bound of the differential of duty i in disjunct j such that $p(a_1,a_2)$ returns true if action a_1 is ethically preferable to action a_2. Therefore, a principle is:

$$p(a_1,a_2) \leftarrow \begin{array}{c} \Delta d_1 \geq v_{1,1} \quad \wedge \quad \cdots \quad \wedge \quad \Delta d_n \geq v_{n,1} \\ \vee \\ \vdots \\ \vee \\ \Delta d_1 \geq v_{n,1} \quad \wedge \quad \cdots \quad \wedge \quad \Delta d_n \geq v_{n,m} \end{array}$$

To conclude this description, here is an example of an ethical dilemma: A health care worker has recommended a particular treatment for her competent adult patient and the patient has rejected that treatment option. Should the health care worker try again to change the patient's mind or accept the patient's decision as final? This dilemma involves the duties of beneficence, non-maleficence, and respect for autonomy and a principle discovered that correctly (as per a consensus of ethicists) balanced these duties in all cases represented. The discovered principle was:

$$p(try\ again, accept) \leftarrow \begin{array}{r} \Delta max\ respect\ for\ autonomy \geq 3 \quad \vee \\ \Delta min\ harm \geq 1 \wedge \Delta max\ respect\ for\ autonomy \geq -2 \quad \vee \\ \Delta max\ benefit \geq 3 \wedge \Delta max\ respect\ for\ autonomy \geq -2 \quad \vee \\ \Delta min\ harm \geq -1 \wedge \Delta max\ benefit \geq -3 \wedge \Delta max\ respect\ for\ autonomy \geq -1 \end{array}$$

A healthcare worker should challenge a patient's decision if it is not fully autonomous and there's either any violation of non-maleficence or a severe violation of beneficence.

Ethical Advisor Systems with Logic Programming In their previous research, the authors of GenEth have also used inductive logic for ethics to build advisor systems.

Jeremy is an advisor system [6] based on action-based ethical theory that provide guidance in ethical decision-making according to Bentham's Hedonistic Act Utilitarianism [22]. The moral decision is made in a straightforward manner. For each possible decision d, there are three components to consider with respect to each person p affected: the intensity of pleasure/displeasure I_p; the duration of the pleasure/displeasureD_p; and the probability that this pleasure/displeasure will occur P_p. This is used to compute the total net pleasure for each decision, the right being the one giving the highest total net pleasure [78]:

$$total_d = \sum_{p \in Person} (I_p \times D_p \times P_p)$$

In the domain of biomedicine, based on prima facie duty theory [74] from biomedical ethics:

- MedEthEx: is dedicated to give advice for dilemmas in biomedical fields [78]. MedEthEx (Medical Ethics Expert) is an implementation of Beauchamp's and Childress' Principles of Biomedical Ethics [21] that harnesses machine learning techniques to abstract decision principles from cases in a particular type of dilemma with conflicting prima facie duties and uses these principles to determine the correct course of action in similar and new cases. MedEthEx helps determine the best course of action in a biomedical ethical dilemma. This approach can be used in the implementation of other such systems that may be based upon different sets of ethical duties and applicable to different domains [9].
- EthEl (ETHical ELdercare system): a medication-reminder system for the elderly and as a notifier to an overseer if the patient refuses to take the medication [7]. EthEl has been implemented in a real robot, the Nao robot, being capable to find and walk toward a patient who needs to be reminded of medication, to bring the medication to the patient, to engage in a natural-language exchange, and to notify an overseer by email when necessary [5].

Addressed Ethical Issue: How to solve an ethical dilemma?

4.3.3 Modeling Morality Computationally with Logic Programming

The implementation techniques for logic programming described here are introduced in [78] and [77]. The authors investigate the potential of logic programming, shortly LP, to model morality aspects studied in philosophy and psychology, by developing an LP-based system with features needed in modeling moral settings, putting emphasis on modeling the following morality aspects:

- dual-process model, reactive and deliberative, in moral judgments
- justification of moral judgments by contractualism
- intention in moral permissibility

To do so, they use the benefits of tabling features to in LP, that is abduction[2] and logic program updates, as a basis into which other reasoning facets are integrated. The implementation techniques presented are:

- Tabled Abduction (TABDUAL): employ tabling mechanisms in logic programs in order to reuse priorly obtained abductive solutions, from one abductive context to another.
- Restricted Evolving Logic Programs (EVOLP/R), adapted from EVOLP by restricting updates at first to fluents (condition that can change over time) only.

Logic Programming Implementation Techniques The tabled logic programming paradigm, is supported by a number of Prolog systems, to different extent. Tabling affords solutions reuse, rather than recomputing them, by keeping in tables subgoals and their answers obtained by query evaluation. The techniques are realized in XSB Prolog, [94] with features such as tabling over default negation, incremental tabling, answer subsumption, call subsumption, and threads with shared tables.

Tabled Abduction, TABDUAL, employs tabling mechanisms in logic programs in order to reuse priorly obtained abductive solutions, from one abductive context to another. It is realized via a program transformation of abductive normal logic programs. Abduction is subsequently enacted on the transformed program. The core transformation of TABDUAL consists of an innovative re-uptake of prior abductive solution entries in tabled predicates and relies on the dual transformation [3], which allows to more efficiently handle the problem of abduction under negative goals, by introducing their positive dual counterparts. In TABDUAL, the dual transformation is refined, to allow it dealing with such programs. The first refinement helps ground (dualized) negative subgoals. The second one allows dealing with non-ground negative goals.

Restricted Evolving Logic Programs, EVOLP/R, is the language described in [77], adapted from that of Evolving Logic Programs (EVOLP) [2], by restricting updates at first to fluents only. More precisely, every fluent is accompanied by its fluent complement. Retraction of the fluent is thus achieved by asserting its complement at the next timestamp, which renders the fluent supervened by its complement at later time; thereby making the fluent false. Nevertheless, it allows paraconsistency, i.e., both the fluent and its complement may hold at the same timestamp, to be dealt with by the user as desired, e.g., with integrity constraints or preferences. In order to update the program with rules, special fluents (termed rule name fluents) are introduced to identify rules uniquely. Such a fluent is placed in the body of a rule, allowing to turn the rule on and off, cf. Poole's "naming device" [72]; this being achieved by asserting or retracting the rule name fluent. The restriction thus requires that all rules be known at the start.

Addressed Ethical Issue: How to model morality?

4.4 Other Approaches to Modeling Ethical Issues

This section describes other models related to ethics, which could not be as neatly categorized in the previous sections:

[2] Abduction is defined as follows: given a logical theory T representing the expert knowledge and a formula Q representing an observation on the problem domain, abductive inference searches for an explanation formula $\mathscr{E}$ such that $\mathscr{E}$ is satisfiable w.r.t. T and it holds that $T \models \mathscr{E} \rightarrow Q$. $\exists(\mathscr{E})$ should be satisfiable w.r.t. T, if $\mathscr{E}$ contains free variables. Generally, if Q and $\mathscr{E}$ contain free variables: $T \models \forall(\mathscr{E} \rightarrow Q)$. Reference: [35]

4.4.1 Simulating Human behaviors in Agent Societies

The model and simulation described here is introduced in [75]. The models is based on sociological research that explores humans' cooperative prosocial behavior, a conceivably non-rational process. It simulates agents that behave like humans. It results in simulated interactions between the human-like agents and a variety of purely rational agents. The simulated scenario is composed of two different game theory games. The first is the Dictator game, which is derivative of the ultimatum game, and consists of a dictator and a receiver. The dictator agent is given a set of resources for which it must choose an amount to donate to a passive receiver agent. It gives this amount to the receiver and keeps the remainder for itself. The second is the Indirect Reciprocity Game, which consists of a dictator and an indirect reciprocator. The game begins with the premise that a dictator game has already occurred. An independent member of the society (the indirect reciprocator) is then asked to indirectly reciprocate the original dictator's behavior from the dictator game. The indirect reciprocator is given a set of resource units, and then told the percentage of the original dictator's resources that were donated to the receiver in the dictator game, as well as situation of the donation (public or private). The indirect reciprocator decides how much of its resources to reciprocate to the dictator.

Models Agents can have the following characterizations: altruists acting privately, altruists acting publicly, egoists acting privately, and egoists acting publicly. In the dictator game agents are modeled as follows:

- Altruists in a private situation donated a mean of 40%;
- Altruists in a public situation donated a mean of 51%;
- Egoists in a private situation donated a mean of 22%;
- Egoists in a public situation donated a mean of 46%

In the indirect reciprocity game agents are modeled as follows:

- An altruist indirectly reciprocating to a dictator that donated in private would reciprocate with an equal proportion of its resources. If the dictator gives 50% of his resources, then the indirect reciprocator gives 50% of his resources.
- An altruist indirectly reciprocating to a dictator that donated in public would match 90% of the percentage the dictator donated to the receiver. If the dictator gives 50%, then the indirect reciprocator gives 45%.
- An egoist indirectly reciprocating to a dictator that donated in private would match 86% of the percentage the dictator donated to the receiver. If the dictator gives 50%, then the indirect reciprocator gives 43%.
- An egoist indirectly reciprocating to a dictator that donated in public would match 64% of the percentage the dictator donated to the receiver. If the dictator gives 50%, then the indirect reciprocator gives 32%

The rational agents, defined in [84] as follows:

- philanthropic agent: is a perpetually cooperative agent. It will always donate 50% of its resources.
- selfish agent: accepts any donations made by others, but never donating anything.
- reciprocative agent: assesses its indebtedness to another agent in its consideration of how much to donate to that agent. This agent will periodically contribute to an agent to which it is not indebted.

Each of these agent types are pitted against each other in both the dictator and indirect reciprocity games. In the Dictator Game, agents of each contending agent type are paired. Dictatorship is then randomly assigned. The dictator is given 8 resource units of which it decides how much to give to the receiver. The transfer of resource units is made. Agents now swap roles so that the receiver becomes the dictator, and the game is played again, ensuring an equal representation of dictator agents from both agent types.

In the Indirect Reciprocator Game, agents of each contending agent type are paired. Indirect reciprocatorship is then randomly assigned. The other agent becomes the dictator. The dictator fabricates a round of the dictator game to produce the amount of resources that the dictator would give to the receiver, but no resources are actually disseminated in this step. The indirect reciprocator receives 9 resource units of which it decides how much to reciprocate to the dictator. The indirect reciprocator gives this amount to the dictator, keeping the remainder for itself. The indirect reciprocator and the dictator swap roles, and the game is played again.

The results show that the amount of resources acquired during each simulation iteration is averaged for each competing agent type and then accrued over many iterations. The rates at which an agent type accumulates resources as compared to its competitor are calculated. The differences between the rates of competing agent types will serve as a metric for characterizing the relative success of one agent type over another. The results for each simulated competition between agent types are shown as the difference between the accumulated resources of one agent type versus the accumulated resources of the contending agent type. These values are identified for all agent type pairs in both the dictator and indirect reciprocator games.

There are two possible interactions indicated by differences in rates of accumulated wealth.

1. Difference is zero. This means that both agent types are gaining and losing resources at the same rate. Neither agent type is benefiting over the other.
2. Difference is not zero. The playing field is not equal between these two agent types. One of the agents is making larger donations to its opponent than it is receiving from its opponent. Such an agent has a greater prosocial tendency. The opponent, on the other hand, is exploiting the agent's prosociality.

The rational agent against which all human agents fare best in the dictator game is the philanthropic agent. Selfish agents perform very well when competing against the human-like agents. The reciprocative rational agent produces nearly balanced resource distribution for all agent types, human and rational, in the dictator and indirect reciprocity games.

Addressed Ethical Issue: How to describe how individuals should behave?

4.4.2 A Simulation of the Argument from Disagreement

The model and simulation described here are introduced in [46]. Argument from Disagreement: widespread and persistent disagreement on ethical issues indicates that our moral opinions are not influenced by any moral facts, either because no such facts exist or because they are epistemically inaccessible or inefficacious for some other reason. This argument is modeled in two different ways, a basic model and an extended model. The basic model is a generalized version of opinion dynamics and bounded confidence models, where opinions can vary continuously in an interval of real numbers, which correspond to the agent's opinion about the degree of moral praiseworthiness of the action under consideration. The extended model is an extension of basic model where persistent moral disagreement is

possible only if the confidence interval is sufficiently small, it assumes the confidence interval is sufficiently small to enable persistent disagreement. The authors show how the Argument from Disagreement can be modeled in a computer simulation, and the outcome seems to lend some additional support to the rejection of non-sceptic moral realism, at least under a wide range of empirical assumptions about how moral opinions evolve over time. [46]

Models The basic model extends generalizes the models of opinion dynamics and bounded confidence in the following ways: the Hegselmann and Krause assume that opinions can vary continuously in an interval of real numbers $[0, 1]$. This is generalized by giving the numbers some intuitive meaning, it is assumed that they correspond to the agent's opinion about the degree of moral praiseworthiness of the action under consideration. In the Hegselmann–Krause model, the outcome of the simulation is determined solely by each agent's willingness to adjust their opinion to similar opinions held by others. If each agent's opinion at each round is influenced only by the view he held the previous round and other views that are within the confidence interval, i.e., views that are sufficiently close to one's own for being worth considering, then consensus will arise just in case the distance between two neighboring opinions is never greater than the confidence interval. The conclusion is that agents will eventually reach agreement on the true opinion, with the exception for the case of an agent whose opinion is not affected at all by the truth, and whose initial opinion is too far apart from that of others, and is thus not affected by those opinions either.

The extended model focuses on investigating whether persistent disagreement is compatible with the assumption that moral opinions are affected by moral facts that are somehow perceived or get acquainted with. In the Hegselmann–Krause model, persistent moral disagreement is possible only if the confidence interval is sufficiently small. Otherwise, consensus will quickly be reached. This is generalized by assuming that the confidence interval is sufficiently small to enable persistent disagreement; and that individuals are even more conservative than before; each individual takes his own view to count for several hundred times as much as his moral peer. The algorithm for this model follows these steps:

1. initialize all variables based on parameters
2. increase round
3. set conservatism
4. set peers
5. set authorities
6. update age
7. update new views of each peer
8. shifts
9. successors
10. terminate if number of total rounds is reached

Findings add support to the rejection of non-sceptic moral realism, at least under a wide range of empirical assumptions about how moral opinions evolve over time. The methodology used is useful for moral philosophers wishing to discuss the meta-ethical significance of moral disagreement. The authors' conclusion is that computer simulations provide a new tool for assessing meta-ethical debates about moral disagreement.

Addressed Ethical Issue: Argument from disagreement

4.4.3 Computational Models of Ethical Reasoning

The ethical reasoning programs described here are introduced in [58]. How can machines support or replace humans in performing ethical reasoning? To answer this question, two ethical reasoning programs are presented:

- Truth-teller: compares pairs of truth-telling cases. It is designed to accept a pair of ethical dilemmas and describe the salient similarities and differences between the cases from both an ethical and pragmatic perspective.
- SIROCCO: retrieves relevant past cases and principles when presented with an ethical dilemma. It is constructed to accept a single ethical dilemma and retrieve other cases and ethical principles that may be relevant to the new case.

The authors mention that ethical reasoning is based on abstract principles that cannot be easily applied in formal, deductive fashion and there is no universal agreement on which ethical theory or approach is the best. The Truth-Teller and SIROCCO projects show that ethical reasoning has a fundamentally different character than reasoning in more formalized domains. These projects also show the difficulty in imbuing a computer program with the sort of flexible intelligence required to perform ethical analysis. The authors propose that computer programs should only act as aids in ethical reasoning. [58]

Ethical Reasoning Programs Ethical reasoning programs focused on reasoning from cases, implementing aspects of the ethical approach known as casuistry. Casuistry is a form of ethical reasoning in which decisions are made by comparing a problem to paradigmatic, real, or hypothetical cases [51].

Truth Teller compares pairs of cases presenting ethical dilemmas about whether or not to tell the truth [11] [10]. The program is intended as a first step in implementing a computational model of casuistic reasoning. The Truth-Teller program marshals ethically relevant similarities and differences between two given cases from the perspective of the "truth teller", i.e., the person faced with the dilemma, and reports them to the user. In particular, it points out reasons for telling the truth, or not, that apply to both cases; apply more strongly in one case than another; or apply to only one case. This program compares pairs of cases given to it as input by aligning and com- paring the reasons that support telling the truth or not in each case. More specifically, Truth-Teller's comparison method comprises four phases of analysis:

1. Alignment: build a mapping between the reasons in the two cases, that is, indicate the reasons that are the same and different across the two representations
2. Qualification: identify special relationships amongst actors, actions, and reasons that augment or diminish the importance of the reasons, for example, telling the truth to a family member is typically more important than telling the truth to a stranger
3. Marshaling: select particular similar or differentiating reasons to emphasize in presenting an argument that one case is as strong as or stronger than the other with respect to a conclusion; the cases are only weakly comparable; or the cases are not comparable at all
4. Interpretation: generate prose that accurately presents the marshaled information so that a nontechnical human user can understand it.

SIROCCO is implemented to bridge the gap between general principles and concrete facts of cases. The program emulates the way an ethical review board within a professional engineering organization (the National Society of Professional Engineers – NSPE) decides cases by referring to, and balancing between, ethical codes and past cases [68].

Its goal is, given a new case to analyze, to provide the basic information with which a human reasoner, for instance a member of the NSPE review board, could answer an ethical question and then build an argument or rationale for that conclusion [59]. This program accepts input, or target, cases in a detailed case-representation language called the Engineering Transcription Language (ETL). SIROCCO's language represents the actions and events of a scenario as a Fact Chronology of individual sentences, i.e., Facts. A predefined ontology of Actor, Object, Fact Primitive, and Time Qualifier types are used in the representation. At least one Fact in the Fact Chronology is designated as the Questioned Fact; this is the action or event corresponding to the ethical question raised in the scenario. SIROCCO utilizes knowledge of past case analyzes, including past retrieval of principles and cases, and the way these knowledge elements were utilized in the past analyses to support its retrieval and analysis in the new target case. The program employs a two-stage graph-mapping algorithm to retrieve cases and codes:

1. Stage 1 performs a "surface match" by retrieving all source cases – the cases in the program's database, represented in an extended version of ETL (EETL), totaling more than four hundred – that share any fact with the target case. It computes a score for all retrieved cases based on fact matching between the target case and each source case, and outputs a list of candidate source cases ranked by scores.
2. Stage 2 using A* search it attempts a structural mapping between the target case and each of the N top-ranking candidate source cases from Stage 1. SIROCCO takes temporal relations and abstract matches into account in this search. The top-rated structural mappings uncovered by the A* search are organized and displayed by a module called the Analyzer.

When comparing Truth-Teller and SIROCCO, the former is more useful in helping users compare cases and recognize important similarities and differences between the cases. The latter is more useful for collecting a variety of relevant information, principles, cases, and additional information that a user should consider in evaluating a new ethical dilemma. Whereas Truth-Teller has a clear advantage in comparing cases and explaining those comparisons, it ignores the problem of how potentially "comparable" cases are identified in the first place. Truth-Teller compares any pair of cases it is provided, no matter how different they may be. SIROCCO uses a retrieval algorithm to determine which cases are most likely to be relevant to a given target case and thus worth comparing. SIROCCO's representational approach is more sophisticated and general than Truth-Teller's. SIROCCO's case comparisons are not nearly as precise and issue-oriented as Truth-Teller's. Both the Truth-Teller and SIROCCO projects are focused and rely heavily on a knowledge representation of ethics. Truth-Teller has a rich representation of truth-telling dilemmas. SIROCCO has a deep representation of engineering ethics principles and engineering scenarios, but no knowledge of more general ethical problem solving.

Addressed Ethical Issue: How can machines support or replace humans in performing ethical reasoning?

CHAPTER 5

Discussion

This chapter focuses on the discussion of the different approaches so far. It additionally describes the relation between ethics and simulation from a philosophical point of view, and concludes with mentions of alternative uses of simulation that relate to ethics.

5.1 Comparison of the Different Approaches

This section discusses the research presented so far, by investigating common properties in the different models and mentioning exceptions.

Most simulations that use agent-based modeling are similar in that they use an environment composed of a two-dimensional world, mostly a grid or network. A few of these models are created in or extend existing NetLogo models, thereby offering an accessible and easily intractable model.

The conditions for the investigated emerging behavior in the model are based on empirical data or theoretical research in the investigated field, in one of the models these conditions are identified using the trace validation technique, namely in the model presented in the section 4.1.8 Similarly, the model architecture or specification is based on theoretical and empirical literature. In the exceptional case of the model presented in the section 4.1.11 evidence-based modeling strategy is used, as defined in [40]. The authors of the paper that describes this model also offer detailed motivations of the model assumptions from the empirical evidence.

The parameter set input to the models for the simulation is based on studies in the research fields the model investigates, or it is based on statistical results, or estimated by the authors, based on their research experience. This parameter set mostly include properties of the modeling environment and the agent. Often they include thresholds, which influence the environment, or the agent behavior.

Agent properties and characteristics are based on studies in the investigated research field, or in the particular case of 4.1.10 in statistical analysis of questions and responses from the International Social Survey Programme [33]. This model also uses structural equation modeling [38] to organize the relationships between specific agent characteristics. The model described in 4.1.1 also offers a detailed analysis of the social network formed by the agents.

The decision making process agents undertake differs in most of the ABM simulations. In some, agents base their decisions on thresholds related to their properties and attributes, resulting in changes in these properties. In other simulations, agents based their decisions according to probability distributions, which themselves are based on empirical data, research, qualitative studies, defined by the authors based on their experience, or simply random. In other simulations, agents based their decisions according specific norms and rules. In other simulations, agents based their decisions according decision trees, every possible action is strictly defined and the agent follows a specific path for each scenario. In some cases these different approaches are combined, for example based on a specific threshold, the agent might act on a behavior based on a norm or a decision tree, as is the case in 4.1.8. Some models use cognitive architectures ti improve the decision making process. Namely, the BDI architecture in 4.3.1 and the PECS architecture in 4.1.12. In the simulations that utilize game theory concepts, the decision making is based on the game theory strategy that is being investigated. In the particular case of the model presented in 4.3.1 agents adapt the decision they can make to realize a judgment based on the information available to them.

The execution and results of the simulations are verified in different ways. Many of the simulations are executed with the tested behavior enabled and disabled, allowing the investigation of the impact of this behavior. Additionally, some simulations compare different groups, that is, they have a test group and a control group. In some models, simulation outputs are compared to empirical data, collected in the investigated research field or are analyzed, for example utilizing an exploratory result analysis, as in the model described in 4.1.1. The validity of the model is demonstrated with a numerical example in 4.2.1, and the evaluation of the codified principles is executed with a specific Turing Test, the "ethical" Turing Test, by the authors of the model described in 4.3.2. Some models are validated by comparing them to other methods. The model described in 4.1.9 is compared to an SDM model with s similar simulation goal. The model described in 4.1.10 instead is evaluated against a baseline approach based entirely on historical data, and a statistical approach that uses linear regression modeling.

The model described in 4.1.6 comes with useful flow chats, which clearly illustrates each possible action agents can undertake and each possible outcome.

5.2 Ethics and Simulation

This section describes the relation of modeling and simulation in ethics, and their implication, and mentions solutions that suggest the ethical considerations.

Modeling and simulation professionals are typically focused on explaining things and solving problems, on figuring out how stuff works and figuring out how to do stuff better [89].

Modeling is the task-driven, purposeful simplification and abstraction of a perception of reality that is shaped by physical, ethical and cognitive constraints [98, p. 47]. Modeling is purposeful, it is driven by human purpose. Models have intention which is explicit, if implicit, it can be easily identified from code. Simulators should attend to models ontological, epistemological assumptions and implications and render explicit the teleology at work in their "purposeful abstractions" [97, p. 11]. Models are based on perception of reality, therefore computer scientists have the moral responsibility to embrace humility when making claims about their models or the outcomes of their simulation experiments [89].

Models are "interpreted structures" [99, p. 39], therefore modelers should be aware of the function of their own "construals" [1] as they interpret these structures. Subjective in-

[1] A construal is a person's perception and interpretation of attributes and behavior of the self or of others. Reference: [12]

terpretations, consturals and perceptions are always and already wrapped up within the power structures within which modelers are socially entangled [89]. To avoid this bias, the validity of the model should not be influenced by the modeler's validation, self-interest, socio-economic status, and relationship dynamics [48, p. 99-100].

Sterner [91] views simulation as a way of carrying out more complex and quantitative thought experiments, however he refuses Mascaro, Korb, Nicholson and Woodberry's argument that simulations of ethical decisions carry the same epistemological weight as experiments with human subjects, without the downside [56]. This is motivated by the fact that, simulation validation is based on experimental or observational data, rendering the simulations epistemologically dependent on empirical methods in a way the empirical methods are not necessarily dependent on simulations. He additionally refutes their methods, as he sees in them the potential to test the coherence or hypothetical plausibility of certain claims within utilitarianism, but none in meta-ethical issues. He suggests that a better approach to tackling ethical issues, would be to pick several simple but important problems for which new results could be generated. The documentation of these cases and the validation of the method would legitimately justify its results. Sterner, criticizes the position that the abstract simplicity of formal models gives them universal scope, saying that their abstractness may render the models applicable to nothing, mentioning scientists working to validate simulations of protein folding, who have spent time and effort overcoming problems in matching models to processes, still acknowledging that important difficulties remain. The language used to make claims about simulation results should be carefully hedged to reflect the distance between the model and the real process, especially for the controversial simulation topics in *Evolving Ethics: The New Science of Good and Evil* [56] [91]. Models for simulations that tackle ethical issues should be specified carefully, taking good consideration of what is being simulated, ensuring that they are not abstracted too far away from what is being represented and validating the execution and results properly.

Shults, Wildman, and Dignum collaborate in creating a framework for ethical analysis of the practice of computer modeling and simulation. They investigate the question: Is the purpose of a given model "good," and if so, for whom? By whose standards? So they propose a meta-ethical framework for exploring the ethics of simulation, which contains philosophical, scientific and practical meta-ethics components, and guidelines for each component. This framework guides the specification or analysis of models that simulate human behaviors within artificial societies [88].

Diallo, Schults and Wildman suggest that incorporating morally salient dimensions of a culture is critically important for producing relevant and accurate evaluations of social policy when using multi-agent artificial intelligence models and simulations [36].

Barlow suggests that both those who build and use simulations should be prepared to exercise an inherent moral responsibility for their work. The builders should additionally be prepared to disclose information about the simulation to the users so that they can interpret the results with references to the output veracity [16]. Disclosure is also a key issue when considering the ethical implications of simplification, which is necessary or modeling and simulation [17].

The Society for Modeling and Simulation sponsored a task group to propose a code of ethics the modeling and simulation community, that observes who is responsible and to whom) [70], which emphasizes the importance of issues such as personal development, professional competence, and commitment to promoting reliable and credible use of modeling and simulation [69]. The code of ethics also addresses the following themes [88]:

- treating employees, clients, users, colleagues, and employers fairly
- endeavoring to seek, utilize, and provide critical professional review
- cautioning against accepting simulation results when there is insufficient evidence of thorough validation and verification
- supporting studies that do not harm humans or the environment
- giving full acknowledgment to the contributions of others

This has been widely accepted and adopted by numerous modeling and simulation societies and organizations, e.g. Simulation Interoperability Standards Organization (SISO), Society for Modeling & Simulation International (SCS) [88].

5.3 Other Uses of Simulation in Ethics

This section shortly mentions alternative uses of modeling and simulation in ethics.

The following are also mentioned in [89]. Computer modeling and simulation can be used as an aid for teaching and understanding ethics: Murragara and Wallace describe an ethics course where it is taught, how to construct agent-based models in which the simulated agents are programmed to represent decision making behaviors guided by utilitarianism, Kantianism, and other ethical theories. Students can the experiment in artificial societies to discover how these various strategies play out in e.g. a natural disaster scenario [64]. Similarly, Perry and Robichaud discuss the educational benefits of using simulation to teach normative theory and present simulation design guiding principles, and apply these to a primaries campaign management [71].

Another use of computer modeling and simulation is to train for ethical behavior or moral decision-making, by giving students the opportunity to explore how different moral principles, line utilitarian calculations, rights, virtue ethics, could influence behavior in particular case studies in business-ethics dilemmas, as described by Schumann, Scott and Anderson in [79].

Another still, is to shed a light on moral behaviors within business networks, like marketing exchange relationships. e.g. IPD in which several strategies are tested within the context of various corporate cultures, as described by Hill and Watkins in [50].

Lastly, massively multiplayer online game are simulation environments where individuals interact in different moral and immoral behaviors which influence the players' moral sensibilities [25]. Bainbridge studies social networks and behavior of players in online games to learn about religious and secular systems in the real world [85].

CHAPTER 6

Conclusion

This thesis has offered an overview of the current state of research in computational ethics. It has done so by introducing the research field of computational ethics, defining different computational methods which are used to investigate ethical issues. The methods described were simulations with agent based modeling tackling issues such as altruism, how values and norms affect behavior, how policy making affects unethical behavior, how anxiety affects behavior, the effects of the virtue of temperance on groups of individuals. The description of these models was followed by simulations that consider game theory concepts to investigate the moral status of manipulation, and how moral sentiments affect decision making. The third approach described, illustrates the use of logic programming to explore how artificial agents make ethical judgments, how to solve ethical dilemmas, how to model morality. This description was followed by a discussion that summarized the properties of the different models considered. Next it illustrated the role of computational ethics in the philosophical field of ethics by discussing relation between ethics and simulation, and concluded with a short mention of alternative uses of simulation related to ethics.

In conclusion, the current research in computational ethics tackles different ethical issues and investigates the evolutionary emergence of ethical behavior using different computational approaches. The main approach is simulations of societies of individuals using agent based modeling. Game theory concepts are used to aid in the agents decision making process. Logic programming is used to aid people in making ethical decisions.

Appendix

6.1 Parameters for ABM Simulations in the Representative Example 3.2

PARAMETER	VALUE
fdf	seasonal rate, sinusoid with period of 120 cycles and magnitude of 20 food units; or constant rate, 20 food units per cycle
Actions	Eat, Reproduce, Migrate, Suicide
Observation	age, health, drought condition
world size	15x15
Maximum entities per cell	infinite
Eating neighborhood	1x1
Observation neighborhood	local
Migration rate	0.0005
Migration neighborhood	3x3
Initial number of agents	225
Initial health	30
Agent age limit	50
Health from food	5 health units
Health required for mating	5 health units
Parental investment	10 health units
Action probability mutation	Initial mutation rate: 0.001; Metamutation rate: $N(0, 0.0001)$
Condition value/operator mutation	Initial mutation rate: 0.001; Metamutation rate: $N(0, 0.0001)$

Table 6.1: Parameters of the simulation: 3.2.2 Suicide as an Evolutionarily Stable Strategy [56, p. 138]

PARAMETER	VALUE
rpp	0.9; 0.75; 0.5
Actions and their initial probabilities	Eat: $\frac{1}{3}$, Mate: $\frac{1}{3}$, Rest: $\frac{1}{12}$, Walk: $\frac{1}{12}$, Turn: $\frac{1}{12}$, Rape: $\frac{1}{12}$
Observations	Age, Health, Sex, Local food density, Local population density, Mate requested
Board size	40x40
Maximum entities per cell	1
Neighborhood size	7x7
Initial population	800
Health required for reproduction	200 health units
Parental investment	From 0 to -300 health units
Genome type	Production rules (7 fixed rules)
fdf	Seasonal: period: 60 cycles; magnitude: 60 food units; mean: 130 food units
Action probability mutation	Initial mutation rates: $N(0, N(0.01, 0.001)2)$; Meta-mutation rate: $N(0, 0.001)$
Condition value mutation	Initial mutation rates: $N(0, N(0.01, 0.001)2)$; Meta-mutation rate: $N(0, 0.001)$

Table 6.2: Parameters of the simulation: 3.2.2 Rape and Sexually Dimorphic Behavior [56, p. 187]

	SUCCESS		FAILURE	
	Utility	Health	Utility	Health
WALK	5	-6	0	0
TURN	1	-2	–	–
REST	2	1	0	0
EAT	10	~ 140	-10	-10

Table 6.3: Simulation: 3.2.2 Rape and Sexually Dimorphic Behavior. Utilities and health effects for actions walk, turn, rest and eat. (A dash indicates an impossible outcome.) On births the health effect equals parental investment. The parental investments for mate and rape are matched — for example, when parental investment after a mate is 300 health units, the female investment after a rape is 590 health units. Mating is identical for males and females, including subsequent investments This also applies to the following tables related to this simulation. [56, p. 188]

MATE		
Outcome	Utility	Health
Request accepted, birth	15	[-300,-240,-180,-120,-60]
Request accepted, no birth	15	-16
Request denied	0	0
Cannot find mate	-10	-10

Table 6.4: Simulation: 3.2.2 Rape and Sexually Dimorphic Behavior. Utilities and health effects for action: mate. [56, p. 188]

RAPE – Victim

Outcome	Sex	Utility	Health
Rape, birth	F	-70	[-590,-470,-350,-230,-110]
	M	-70	-10
Rape, no birth	F	-70	-10
	M	-70	-10
Rape attempt prevented	F	-10	-10
	M	-10	-10

Table 6.5: simulation: 3.2.2 Rape and Sexually Dimorphic Behavior. Utilities and health effects for simulation for rape victim. [56, p. 188]

RAPE – Rapist

Outcome	Sex	Utility	Health
Rape, birth	F	5	[-590,-470,-350,-230,-110]
	M	5	-10
Rape, no birth	F	5	-10
	M	5	-10
Rape attempt prevented	F	-15	-60
	M	-15	-60

Table 6.6: Simulation: 3.2.2 Rape and Sexually Dimorphic Behavior. Utilities and health effects for simulation for rapist. [56, p. 188]

PARAMETER	VALUE
Gestation cycles	5
fdf	Constant-rate: 50 food units per cycle; Periodic drought: 50 food units for 40 cycles, 0 food units for 8 cycles
Actions and initial probabilities	Eat: $\frac{4}{9}$, Mate: $\frac{4}{9}$, Abortion: $\frac{1}{9}$
Observations	Health, Global food density, Is gestating?
Board size	25x25
Maximum entities per cell	1
Neighborhood	7x7
Initial population	400
Agent age limit	$N(100, 15^2)$ cycles
Health obtainable from food	$N(140, 10^2)$ health units
Health required for reproduction	200 health units
Genome type	Decision tree
Action probability mutation	Initial mutation rates: $N(0, N(0.01, 0.001^2)^2)$, Meta-mutation rate: $N(0, 0.001^2)$
Branch node value mutation	Initial mutation rates: $N(0, N(0.01, 0.001^2)^2)$, Meta-mutation rate: $N(0, 0.001^2)$
Misc.	Male and Female Sexes; Gestational investment uniform

Table 6.7: Parameters of the simulation: Abortion [56, p. 212]

	SUCCESS		FAILURE	
ACTION	Utility	Health	Utility	Health
EAT	10	*sim*70	-10	-10
ABORTION	0	-40	0	-40

Table 6.8: Simulation: 3.2.2 Abortion. The utilities and health effects associated with the outcomes of actions eat and abortion [56, p. 213]

MATE Action			
Outcome	Sex	Utility	Health
No conception		15	-16
Conception	F	15	-15
+ Total Gestation	F	0	-20
			-150
			-300
+ After Birth	F	0	-20
			-150
			-300
Conception	M	15	-10
Request denied		0	0
Cannot find mate		-10	-10

Table 6.9: Simulation: 3.2.2 Abortion. The utilities and health effects associated with the outcomes of mate actions [56, p. 213]

Bibliography

[1] Anthony Aaby. Computational ethics. *Creative Commons Nathan Abbott Way, Stanford, California 94305, USA, A work in progress; draft as of*, 2005.

[2] José Júlio Alferes, Antonio Brogi, Joao Alexandre Leite, and Luís Moniz Pereira. Evolving logic programs. In *European Workshop on Logics in Artificial Intelligence*, pages 50–62. Springer, 2002.

[3] José Júlio Alferes, Luís Moniz Pereira, and Terrance Swift. Abduction in well-founded semantics and generalized stable models via tabled dual programs. *Theory and Practice of Logic Programming*, 4(4):383–428, 2004.

[4] Sudhir Anand and Amartya Sen. Human development Index: Methodology and Measurement. Human Development Occasional Papers (1992-2007) HDOCPA-1994-02, Human Development Report Office (HDRO), United Nations Development Programme (UNDP), June 1994. URL: `https://ideas.repec.org/p/hdr/hdocpa/hdocpa-1994-02.html`.

[5] Michael Anderson and Susan Anderson. Robot be good. *Scientific American*, 303:72–7, 10 2010. `doi:10.1038/scientificamerican1010-72`.

[6] Michael Anderson, Susan Anderson, and C. Armen. Towards machine ethics: Implementing two action-based ethical theories. *AAAI Fall Symposium - Technical Report*, pages 1–7, 01 2005.

[7] Michael Anderson and Susan Leigh Anderson. Ethel: Toward a principled ethical eldercare robot. *AAAI Fall Symposium: AI in Eldercare: New Solutions to Old Problems*, 2008.

[8] Michael Anderson and Susan Leigh Anderson. Geneth: A general ethical dilemma analyzer. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 253–261. AAAI Press, 2014.

[9] Michael Anderson, Susan Leigh Anderson, and Chris Armen. Medethex: A prototype medical ethics advisor. In *AAAI*, pages 1759–1765. AAAI Press, 2006. URL: `http://dblp.uni-trier.de/db/conf/aaai/aaai2006.html#AndersonAA06`.

[10] Kevin D Ashley. Case-based comparative evaluation in truth-teller. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, volume 17, page 72. Psychology Press, 1995.

[11] Kevin D Ashley and Bruce M McLaren. Reasoning with reasons in case-based comparisons. In *International conference on case-based reasoning*, pages 133–144. Springer, 1995.

[12] American Psychological Association. Construal, apa dictionary of psychology. [Online; accessed 01.04.2022]. URL: https://dictionary.apa.org/construal.

[13] Robert Axelrod. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, 1997. URL: http://www.jstor.org/stable/j.ctt7s951.

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

[15] J. Banks. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. A Wiley-Interscience publication. Wiley, 1998. URL: https://books.google.de/books?id=dMZ1Zj3TBgAC.

[16] John Barlow. Computer simulations, disclosure and duty of care. *Australasian Journal of Information Systems*, 13(2), May 2006. URL: https://journal.acs.org.au/index.php/ajis/article/view/51, doi:10.3127/ajis.v13i2.51.

[17] John Barlow. Simplification: ethical implications for modelling and simulation. In *Proceedings of the 18th world IMACS/MODSIM congress, Cairns, Australia*, pages 432–438. Citeseer, 2009.

[18] Ana Bazzan, Rafael Bordini, and John Campbell. *Evolution of Agents with Moral Sentiments in an Iterated Prisoner's Dilemma Exercise*, volume 5, pages 43–64. Springer US, 07 2011. doi:10.1007/978-1-4615-1107-6_3.

[19] Ana L. C. Bazzan, Rafael H. Bordini, and John A. Campbell. Moral sentiments in multi-agent systems. *Intelligent Agents V—Proceedings of the Fifth International Workshop on Agent Theories, Architectures, and Languages*, pages 113–131, 1999. doi:10.1007/3-540-49057-4_8.

[20] Campbell J. A. Bazzan A., Bordini R. H. Agents with moral sentiments in an iterated prisoner's dilemma exercise. In *In Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents*, November 1997.

[21] T.L. Beauchamp and J.F. Childress. *Principles of Biomedical Ethics*. Oxford University Press, 2009. URL: https://books.google.de/books?id=xg8iwAEACAAJ.

[22] Laurence J. Bentham Jeremy, Lafleur. *An introduction to the principles of morals and legislation*. Hafner Pub. Co., New York, N.Y., 1948.

[23] Daniel Birks, Michael Townsley, and Anna Stewart. Generative Explanations of Crime: Using Simulation to Test Criminological Theory*. *Criminology*, 50(1):221–254, 2012. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-9125.2011.00258.x, doi:10.1111/j.1745-9125.2011.00258.x.

[24] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences*, 99(suppl_3):7280–7287, 2002. URL: https://www.pnas.org/doi/abs/10.1073/pnas.082080899, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.082080899, doi:10.1073/pnas.082080899.

[25] Hope Botterbusch and R.s Talab. Copyright and you: Ethical issues in second life. *TechTrends*, 53:9–12, 01 2009. doi:10.1007/s11528-009-0227-4.

[26] Miles Brundage. Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence*, 26(3):355–372, 2014.

[27] Richard W. Byrne and Andrew Whiten. Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans. *Behavior and Philosophy*, 18(1):73–75, 1990.

[28] Scemama Paul Caglayan Pinar. Computer science in the majors: Agent-based modeling with netlogo, April 2021. PowerPoint presentation; Created by Developer Student Club at W&M, Fall 2020. URL: https://github.com/developerstudentclubwm/cs_majors#workshop-6-agent-based-modeling-with-netlogo.

[29] Christopher A. Chung. *Simulation Modeling Handbook: A Practical Approach*. CRC Press, Inc., USA, 2003.

[30] Julio B. Clempner. A game theory model for manipulation based on machiavellianism: Moral and ethical behavior. *Journal of Artificial Societies and Social Simulation*, 20(2):12, 2017. URL: http://jasss.soc.surrey.ac.uk/20/2/12.html, doi:10.18564/jasss.3301.

[31] Nicolas Cointe, Grégory Bonnet, and Olivier Boissier. Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, page 1106–1114, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems.

[32] Marco Conti, Andrea Passarella, and Fabio Pezzoni. A model for the generation of social network graphs. In *2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, pages 1–6. IEEE, 2011.

[33] J A Davis and R Jowell. Measuring national differences: an introduction to the international social survey programme (ISSP). In R Jowell, S Witherspoon, and & L Brook, editors, *British Social Attitudes: Special International Report*, pages 1–13. Gower, Aldershot, 1989.

[34] Francien Dechesne, Gennaro Di Tosto, Virginia Dignum, and Frank Dignum. No smoking here: Values, norms and culture in multi-agent systems. *Artif. Intell. Law*, 21(1):79–107, mar 2013. doi:10.1007/s10506-012-9128-5.

[35] Marc Denecker and Antonis Kakas. Abduction in logic programming. In *Computational logic: Logic programming and beyond*, pages 402–436. Springer, 2002.

[36] Saikou Y. Diallo, F. LeRon Shults, and Wesley J. Wildman. Minding morality: ethical artificial societies for public policy modeling. *AI & society*, pages 1–9, Aug 2020. 32836907[pmid], PMC7411344[pmcid], 1028[PII]. doi:10.1007/s00146-020-01028-5.

[37] R. I. M. Dunbar and Susanne Shultz. Evolution in the social brain. *Science*, 317(5843):1344–1347, 2007. URL: https://www.science.org/doi/abs/10.1126/science.1145463, arXiv:https://www.science.org/doi/pdf/10.1126/science.1145463, doi:10.1126/science.1145463.

[38] Otis Dudley Duncan. *Introduction to structural equation models*. Elsevier, 2014.

[39] Abeer Dyoub, Stefania Costantini, and Francesca A. Lisi. Logic programming and machine ethics. *Electronic Proceedings in Theoretical Computer Science*, 325:6–17, Sep 2020. URL: http://dx.doi.org/10.4204/EPTCS.325.6, doi:10.4204/eptcs.325.6.

[40] Bruce Edmonds and Scott Moss. From kiss to kids: ¡i¿an 'anti-simplistic' modelling approach¡/i¿. In *Proceedings of the 2004 International Conference on Multi-Agent and Multi-Agent-Based Simulation*, MABS'04, page 130–144, Berlin, Heidelberg, 2004. Springer-Verlag. doi:10.1007/978-3-540-32243-6_11.

[41] Joshua M. Epstein. Why model? *Journal of Artificial Societies and Social Simulation*, 11(4), October 2008. Copyright: Copyright 2008 Elsevier B.V., All rights reserved.

[42] Joshua M. Epstein and Robert L. Axtell. *Growing Artificial Societies: Social Science from the Bottom Up*. The MIT Press, 10 1996. doi:10.7551/mitpress/3374.001.0001.

[43] Nigel Gilbert and Klaus G Troitzsch. *Simulation for the Social Scientist*. Open University Press, USA, 2005.

[44] Angel Gómez, Matthew L Brooks, Michael D Buhrmester, Alexandra Vázquez, Jolanda Jetten, and William B Swann Jr. On the nature of identity fusion: insights into the construct and a new measure. *Journal of personality and social psychology*, 100(5):918, 2011.

[45] Ross Gore, Carlos Lemos, F. LeRon Shults, and Wesley J. Wildman. Forecasting changes in religiosity and existential security with an agent-based model. *Journal of Artificial Societies and Social Simulation*, 21(1), January 2018. doi:10.18564/jasss.3596.

[46] Johan E. Gustafsson and Martin Peterson. A computer simulation of the argument from disagreement. *Synthese*, 184(3):387–405, 2012. URL: http://www.jstor.org/stable/41411200.

[47] Keith Hankins and Peter Vanderschraaf. Game Theory and Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

[48] Brian L. Heath and Ross A. Jackson. Ontological implications of modeling and simulation in postmodernity. In *Ontology, Epistemology, and Teleology for Modeling and Simulation*, 2013.

[49] Joseph Henrich. The evolution of costly displays, cooperation and religion: credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior*, 30(4):244–260, 2009. URL: https://www.sciencedirect.com/science/article/pii/S1090513809000245, doi:https://doi.org/10.1016/j.evolhumbehav.2009.03.005.

[50] Ronald Paul Hill and Alison Watkins. A simulation of moral behavior within marketing exchange relationships. *Journal of the Academy of Marketing Science*, 35(3):417–429, Sep 2007. doi:10.1007/s11747-007-0025-5.

[51] Albert R Jonsen, Stephen Toulmin, and Stephen Edelston Toulmin. *The abuse of casuistry: A history of moral reasoning*. Univ of California Press, 1988.

[52] Stefan Körner. On the origin of altruism – an agent-based social evolutionary simulation. Masterarbeit/master thesis, Institute of Computer Science, LMU, Munich, 2022. URL: http://www.pms.ifi.lmu.de/publikationen/#DA_Stefan.Koerner.

[53] Jeremiah Lasquety-Reyes. Computer Simulations of Ethics: the Applicability of Agent-Based Modeling for Ethical Theories. *European Journal of Formal Sciences and Engineering*, 1:18, July 2018. doi:10.26417/ejfe.v1i2.p18-28.

[54] Jeremiah A. Lasquety-Reyes. Towards Computer Simulations of Virtue Ethics. *Open Philosophy*, 2(1):399–413, 2019. doi:10.1515/opphil-2019-0029.

[55] Hector J Levesque. Knowledge representation and reasoning. *Annual review of computer science*, 1(1):255–287, 1986.

[56] Steven Mascaro, Kevin Burt Korb, Ann Elizabeth Nicholson, and Owen Grant Woodberry. *Evolving Ethics: The New Science of Good and Evil*. Imprint Academic, 1 edition, December 2010. URL: https://www.ebook.de/de/product/10403988/steven_mascaro_kevin_b_korb_ann_e_nicholson_evolving_ethics_the_new_science_of_good_and_evil.html.

[57] Francis T. McAndrew. *Costly Signaling Theory*, pages 1–8. Springer International Publishing, Cham, 2018. doi:10.1007/978-3-319-16999-6_3483-1.

[58] B.M. McLaren. Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems*, 21(4):29–37, 2006. doi:10.1109/MIS.2006.67.

[59] Bruce M McLaren. Extensionally defining principles and cases in ethics: An ai model. *Artificial Intelligence*, 150(1-2):145–181, 2003.

[60] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[61] Rijk Mercuur, Virginia Dignum, and Catholijn Jonker. The value of values and norms in social simulation. *Journal of Artificial Societies and Social Simulation*, 22(1):1–9, 2019. doi:10.18564/jasss.3929.

[62] James H Moor. Is ethics computable? *Metaphilosophy*, 26(1/2):1–21, 1995.

[63] Ioan Muntean and Don Howard. A minimalist model of the artificial autonomous moral agent (aama). In *SSS-16 Symposium Technical Reports. Association for the Advancement of Artificial Intelligence*. AAAI, 2016.

[64] Ruth I. Murrugarra and William A. Wallace. Agent-based simulation for teaching ethics. In *Proceedings of the 2017 Winter Simulation Conference*, WSC '17. IEEE Press, 2017.

[65] Martin Neumann. The escalation of ethnonationalist radicalization: Simulating the effectiveness of nationalist ideologies. *Social Science Computer Review*, 32(3):312–333, 2014. arXiv:https://doi.org/10.1177/0894439313511585, doi:10.1177/0894439313511585.

[66] Pippa Norris and Ronald Inglehart. *Sacred and secular: Religion and politics worldwide*. Cambridge University Press, 2011.

[67] National Institute of Biomedical Imaging and Bioengineering. Computational modeling, May 2020. [Online; accessed 23.03.2022]. URL: https://www.nibib.nih.gov/science-education/science-topics/computational-modeling.

[68] National Society of Professional Engineers. The nspe ethics reference guide, May 1996. [Online; accessed 16.02.2022], Alexandria, VA : the National Society of Professional Engineers. URL: https://www.nspe.org/resources/ethics/code-ethics.

[69] TI Oren, Maurice S Elzas, Iva Smit, and Louis G Birta. Code of professional ethics for simulationists. In *Summer computer simulation conference*, pages 434–435. Society for Computer Simulation International; 1998, 2002.

[70] Tuncer Ören. Responsibility, ethics and simulation. *Transactions of The Society for Computer Simulation International*, 17:165–170, 2000.

[71] Tomer J. Perry and Christopher Robichaud. Teaching ethics using simulations: Active learning exercises in political theory. *Journal of Political Science Education*, 16(2):225–242, 2020. arXiv:https://doi.org/10.1080/15512169.2019.1568879, doi:10.1080/15512169.2019.1568879.

[72] David Poole. A logical framework for default reasoning. *Artificial intelligence*, 36(1):27–47, 1988.

[73] Edy Portmann and Sara D'Onofrio. Computational ethics. *HMD Praxis der Wirtschaftsinformatik*, mar 2022. doi:10.1365/s40702-022-00855-y.

[74] W. D. Ross. *The Right and the Good.Some Problems in Ethics*. Clarendon Press, 1930.

[75] Alicia Ruvinsky and Michael N. Huhns. Simulating human behaviors in agent societies. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '08, page 1513–1516, Richland, SC, 2008. International Foundation for Autonomous Agents and Multiagent Systems.

[76] Alicia I. Ruvinsky. Computational ethics. In *Encyclopedia of Information Ethics and Security*, pages 76–82. IGI Global, 2007. doi:10.4018/978-1-59140-987-8.ch012.

[77] Ari Saptawijaya and Luıs Moniz Pereira. Towards practical tabled abduction usable in decision making. In *Intelligent Decision Technologies*, pages 429–438. IOS Press, 2013.

[78] Ari Saptawijaya and Luís Moniz Pereira. Towards modeling morality computationally with logic programming. In Matthew Flatt and Hai-Feng Guo, editors, *Practical Aspects of Declarative Languages*, pages 104–119, Cham, 2014. Springer International Publishing.

[79] Paul L. Schumann, Timothy W. Scott, and Philip H. Anderson. Designing and introducing ethical dilemmas into computer-based business simulations. *Journal of Management Education*, 30(1):195–219, 2006. arXiv:https://doi.org/10.1177/1052562905280844, doi:10.1177/1052562905280844.

[80] Nick Scott, Aaron Hart, James Wilson, Michael Livingston, David Moore, and Paul Dietze. The effects of extended public transport operating hours and venue lockout policies on drinking-related harms in melbourne, australia: Results from simdrink, an agent-based simulation model. *International Journal of Drug Policy*, 32:44–49, jun 2016. URL: https://www.sciencedirect.com/science/article/pii/S0955395916300251, doi:https://doi.org/10.1016/j.drugpo.2016.02.016.

[81] Nick Scott, Michael Livingston, Aaron Hart, James Wilson, David Moore, and Paul Dietze. SimDrink: An Agent-Based NetLogo Model of Young, Heavy Drinkers for Conducting Alcohol Policy Experiments. *Journal of Artificial Societies and Social Simulation*, 19(1):1–10, 2016. URL: https://ideas.repec.org/a/jas/jasssj/2015-60-4.html, doi:10.18564/jasss.2943.

[82] Samuel T. Segun. From machine ethics to computational ethics. *AI and Society*, 36(1):263–276, 2021. doi:10.1007/s00146-020-01010-1.

[83] Stefan Seil and Libuše Hannah Vepřek. Benefits of altruism as a costly practice on-group stability. a simulation-based investi-gation. Proposed: Social Psychological and Personality Science.

[84] Sandip Sen. Reciprocity: a foundational principle for promoting cooperative behavior among self-interested agents. In *Proceedings of the Second International Conference on Multiagent Systems*, volume 315321, 1996.

[85] Daniel B. Shank. eGods: Faith versus Fantasy in Computer Gaming. *Sociology of Religion*, 75(1):175–176, 03 2014. arXiv:https://academic.oup.com/socrel/article-pdf/75/1/175/6873466/sru014.pdf, doi:10.1093/socrel/sru014.

[86] F. LeRon Shults, Justin E. Lane, Wesley J. Wildman, Saikou Diallo, Christopher J. Lynch, and Ross Gore. Modelling terror management theory: computer simulations of the impact of mortality salience on religiosity. *Religion, Brain & Behavior*, 8(1):77–100, 2018. arXiv:https://doi.org/10.1080/2153599X.2016.1238846, doi:10.1080/2153599X.2016.1238846.

[87] F. LeRon Shults, Christopher Lynch, Wesley Wildman, Ross Gore, Justin Lane, and Monica Toft. A generative model of the mutual escalation of anxiety between religious groups. *Journal of Artificial Societies and Social Simulation, The*, 21:7, 09 2018. doi:10.18564/jasss.3840.

[88] F LeRon Shults, Wesley J Wildman, and Virginia Dignum. The ethics of computer modeling and simulation. In *2018 Winter simulation conference (WSC)*, pages 4069–4083. IEEE, 12 2018. doi:10.1109/WSC.2018.8632517.

[89] Fount LeRon Shults and Wesley J. Wildman. Ethics, computer simulation, and the future of humanity. *New Approaches to the Scientific Study of Religion*, 2019.

[90] Cavagnetto Stefano and Gahir Bruce. Game theory - its applications to ethical decision making. *CRIS - Bulletin of the Centre for Research and Interdisciplinary Study*, 2014(1), 2014. URL: https://EconPapers.repec.org/RePEc:vrs:bucris:v:2014:y:2014:i:1:p:19:n:5.

[91] Beckett Sterner. Agent-based computer simulation and ethics. *Metascience*, 21(2):403–407, Jul 2012. doi:10.1007/s11016-012-9660-7.

[92] Rob Stocker. Review: Evolving ethics: The new science of good and evil, 2011. [Journal of Artificial Societies and Social Simulation; Online; accessed 04.02.2022]. URL: https://www.jasss.org/14/3/reviews/3.html.

[93] William B Swann Jr, Ángel Gómez, John F Dovidio, Sonia Hart, and Jolanda Jetten. Dying and killing for one's group: Identity fusion moderates responses to intergroup versions of the trolley problem. *Psychological Science*, 21(8):1176–1183, 2010.

[94] Terrance Swift and David S Warren. Xsb: Extending prolog with tabled logic programming. *Theory and Practice of Logic Programming*, 12(1-2):157–187, 2012.

[95] Henri Tajfel. Social psychology of intergroup relations. *Annual review of psychology*, 33(1):1–39, 1982.

[96] Henri Tajfel, John C Turner, William G Austin, and Stephen Worchel. An integrative theory of intergroup conflict. *Organizational identity: A reader*, 56(65):9780203505984–16, 1979.

[97] A. Tolk. Truth, trust, and turing - implications for modeling and simulation. In *Ontology, Epistemology, and Teleology for Modeling and Simulation*, 2013.

[98] Andreas Tolk. *Code of Ethics*, pages 35–52. John Wiley & Sons, Ltd, 2017. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119288091.ch3, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119288091.ch3, doi:https://doi.org/10.1002/9781119288091.ch3.

[99] Michael Weisberg. *Simulation and similarity using models to understand the world*. Oxford University Press, Oxford, 2015.

[100] Harvey Whitehouse, Brian McQuinn, Michael Buhrmester, and William B Swann. Brothers in arms: Libyan revolutionaries bond like family. *Proceedings of the National Academy of Sciences*, 111(50):17783–17785, 2014.

[101] Wikipedia contributors. Ethics — Wikipedia, the free encyclopedia, 2022. [Online; accessed 28-March-2022]. URL: https://en.wikipedia.org/w/index.php?title=Ethics&oldid=1073798003.

[102] Wesley J. Wildman and Richard Sosis. Stability of groups with costly beliefs and practices. *Journal of Artificial Societies and Social Simulation*, 14(3):6, 2011. doi:10.18564/jasss.1781.

[103] U Wilensky. Netlogo, 1999. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL. URL: http://ccl.northwestern.edu/netlogo/.

[104] Uri Wilensky and William Rand. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. The MIT Press, 2015. URL: http://www.jstor.org/stable/j.ctt17kk851.

[105] Lu Yang and Nigel Gilbert. Getting away from numbers: Using qualitative observation for agent-based modeling. *Advances in complex systems*, 11(02):175–185, 2008.