ISSN 2601-8683 (Print)
ISSN 2601-8675 (Online)
European Journal of
Formal Sciences and Engineering
January - June 2019
Volume 2, Issue 1

Computer Simulations of Ethics: The Applicability of Agent-Based Modeling for Ethical Theories

Jeremiah A. Lasquety-Reyes

Universität Hamburg, Germany

Abstract

I consider the applicability of Agent-Based Modeling (ABM) and computer simulations for ethical theories. Though agent-based modeling is already well established in the social sciences, it has not yet found acceptance in the field of philosophical ethics. Currently, there are only a few works explicitly connecting ethics with agent-based modeling. In this paper, I show that it is possible to build computer simulations of ethical theories and that there are also potential benefits in doing so: (1) the opportunity for virtual ethical experiments that are impossible to do in real life, and (2) an increased understanding and appreciation of an ethical theory either through the programming implementation or through the visual simulation. In the first part of the paper, I mention some social science simulations with ethical import that could encourage ethicists to work with ABM. Second, I list the few pioneering works that attempt to combine computer simulation with philosophical ethics, the most prominent being *Evolving Ethics: The New Science of Good and Evil* (2010) by Mascaro et al. Third, I give pointers for the computer simulation of the most prominent ethical theories: deontological ethics, utilitarianism, feminist care ethics, and virtue ethics. In the final part, I consider the potential of using an existing reference model for the simulation of human behavior, the PECS model, as the foundation for a computer simulation of virtue ethics.

Keywords: Ethics, Computer Simulation, Agent-Based Modeling, Virtue Ethics

Introduction

Agent-Based Modeling (ABM) is an established approach in the computational social sciences. It uses computers to simulate the behavior of "agents." Agents may be molecules, organisms, or other entities, but in the context of computational social science, they normally represent human beings with certain behaviors. The interactions of agents with other agents, or agents with their environment, can lead to different results depending on the assignment of specific conditions and values. They can also lead to unforeseen or surprising results called *emergent* behavior, where a complex property at the macro system level is produced that is not encoded at the individual agent level

(Axelrod, 1997, p. 4; Wilensky & Rand, 2015, 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 (Epstein & Axtell, 1996, p. 4; Gilbert & Troitzsch, 2005, p. 14). They are particularly attractive for social scientists because many social experiments cannot be practically (or ethically) carried out in the real world.

Though many social scientists have embraced ABM as a promising approach to conducting social science research, the same cannot yet be said for ethical and moral philosophers. Many philosophers discuss the ethics of computers and technology, but there are only a few who *use* computers—and in particular, computer simulation—to conduct their research on ethics. Given this situation, this paper explores the potential of computer simulation to be used for the benefit of philosophical ethics.

Social Science Simulations and Ethical Implications

Many of the phenomena that computational social scientists study have ethical implications. For example, the first attempt to apply agent-based modeling to social science was by Thomas Schelling. In his book *Micromotives and Macrobehavior* (1978), he presented a model that showed how housing segregation between races could occur even if no individual wanted it specifically, so long as these individuals had a preference not to be an extreme minority in their neighborhood. In other words, individual preferences and actions resulted in large-scale consequences that none of the individual agents actually intended. This phenomenon could be witnessed in the simulation where two populations of different colors would begin mixed together, but slowly start to form their own homogenous neighborhoods. One could also adjust parameters such as the percentage of tolerance of agents to having different colored neighbors. This results in different kinds of segregation. Though the dynamics explored in the simulation were not of a directly ethical nature (the individual preferences and the resulting segregation were not themselves judged as being right or wrong), it obviously touches on many sensitive ethical issues in the real world such as racial discrimination, unequal opportunities and development in cities, etc.

A more recent trend in social science simulation is the investigation of trust. Doloswala investigates the behavior of peer groups when confronted with the discovery of a liar in their group (Doloswala, 2014). Using proximal space to represent the idea of "shared cognitive space," they simulate how a discovered liar would be ostracized from the group depending on different parameters such as the probability of being discovered, the penalty for lying, and the forgetfulness of agents. According to their simulations, the forgetfulness of agents plays a greater role in shaping groups than the penalty for lying. Meanwhile, Lim et al. investigate the interplay between trust at the level of individuals and the development of collective social moral norms (Lim, Stocker, & Larkin, 2008). They call their resulting model a Computational Model of Ethical Trust (CMET), a two-tier architecture that utilizes both agent-based modeling and artificial neural networks.

Kim (2009) and Tykhonov et al. (2008) investigate trust in the context of supply chains and networks. Kim finds that as a trust relationship between trading partners is prolonged and uncertainties about the trustworthiness of trading partners are diminished, one sees a greater stability in their inventory levels over time. This occurs even without any explicit information sharing among trading partners regarding the status of their own inventories. Meanwhile, Tykhonov et al. employ a human "trust and tracing game" with real-life participants to collect data for trust, deceit, and negotiation behavior, and then use that data to inform a computer agent-based model. They hope that this combined research method will produce a model that is more applicable to real-world trade processes.

In these computer simulations, the researchers touch on many ethical issues. For example, they discuss the importance of being recognized as "trustworthy" by others, the damage and disruption caused by lying and deceit, the benefits of a strong trust relationship, etc. Though none of them refer to any ethical theory, there is no obvious reason why an ethical theory cannot be employed to interpret and engage with their data and results. It is clear that social scientists are able to use computer simulations for studies with ethical import. Can philosophers use computer simulation for ethics itself?

Computer Simulations and Ethical Theories

Robert Axelrod is famous for the computer tournament he organized for the iterated Prisoner's Dilemma[1] where many different strategies were submitted from all over the world and matched with each other in round-robin (Axelrod, 1984). The strategy TIT FOR TAT—which first cooperates then subsequently replicates the other player's previous action—emerged as the winner of the entire two-tier tournament. The tournament was not itself an agent-based model, but similar to one in that "agents" (with respective strategies) interacted with all the other agents in the virtual arena. The project also did not espouse any particular ethical theory, though it did explore real-life examples of the TIT FOR TAT strategy. Eventually, Axelrod attempted to move beyond the two-person format of the Prisoner's Dilemma and explore how cooperation could emerge between many individuals simultaneously. In order to do this, he consciously resorted to agent-based modeling (Axelrod, 1997). With the use of ABM, he was able to explore social phenomena such as the promotion of norms, choosing sides, and the formation of new political groups. He was also able to introduce the evolution of

[1] The Prisoner's Dilemma is an imaginary situation employed in game theory. Two prisoners are accused of a crime and placed in separate cells. If one of them confesses and the other does not, the one who confesses receives only 1 year in prison, while the other who does not confess receives 4 years. If both of them do not confess, they each receive only 2 years. If they both confess, they each receive 3 years. While deciding on what to do, they are unable to communicate with each other. Clearly, the greatest payout is if one confesses while the other does not, and we might expect a prisoner to pursue this action out of self-interest. However, the other prisoner might also have the same mindset, and if they both pursue their self-interest, they would be in a worse situation (3 yrs. each) than if they both keep silent (2 yrs. each). What is the best thing for a prisoner to do? Confess (sometimes called "defect") or not confess ("cooperate")? The Prisoner's Dilemma is a one-shot game where both players make their moves simultaneously. In comparison, the iterated Prisoner's Dilemma allows for many moves and a memory of what transpired in the previous moves.

strategies through genetic algorithms. Though containing much ethical import, these simulations did not refer to any particular ethical theory.

Building on Axelrod's work, Peter Danielson was perhaps the first person to explicitly address a philosophical ethical theory with the use of a computer simulation. He coined the term "artificial morality" for a combination of game theory and artificial intelligence used to develop an ethical theory called *instrumental contractarianism*, which is partially based on the work of David Gauthier (Danielson, 1992, p. 17). He used an Extended Prisoner's Dilemma which involves two sequential moves instead of the two simultaneous ones in the traditional Prisoner's Dilemma. He also eschewed the iteration found in Axelrod. The tournament tested whether the "constrained maximizer" of Gauthier, which cooperates with those who cooperate and defects with the rest, really fares better in every case over "straightforward maximizers" (which includes TIT FOR TAT). His simulations answered in the positive. However, according to him, there is apparently another agent, a so-called "reciprocal co-operator" (which cooperates only when cooperation is necessary and sufficient for the other's cooperation), that fares better than Gauthier's constrained maximizer in a varied population environment.

An unexpected conclusion from Danielson is that his "artificial morality" applies more to formal organizations, firms and machines than to actual people (Danielson, 1992, p. 198). According to him, this is because of the lack of cognitive transparency on the part of human beings as well as the unpredictable and lasting influence of emotions. As he says, "it should not be surprising if traditional human morality fares poorly in terms of rational performance... Artificial Morality may lead us to discover techniques of communication and commitment that are morally effective but unavailable to unaided human beings" (Danielson, 1992, p. 201). Regardless of the controversial conclusion, Danielson's work is noteworthy for being the first to combine a philosophical ethical theory with a computer simulation, though his simulation was a tournament in the style of Axelrod and not an ABM.

Alicia Ruvinsky has called for "the integration of computer simulation and ethics theory... an agent-based simulation mechanism that takes a computational perspective to ethics theory" (Ruvinsky, 2008, p. 76). She uses the term "computational ethics" for this project but this can be confusing since people also use the same term to refer to ethics for AI (i.e. machine ethics) or ethics for computer programmers, so I do not follow her in calling the combination of computer simulation and ethics "computational ethics." Her short article does not give any implementation but only some rough suggestions. According to her interpretation, "an ethic is a moral framework characterized by rights, liberties, and duties, which are parameters in an ethic model" (Ruvinsky, 2008, p. 77). She then claims that ethical theories such as deontological ethics and divine command ethics can be quantified using these parameters. Once such ethical theories are quantified, one can simulate artificial societies where agents can adopt different ethical theories and interact with each other.

These simulations are useful in considering emergent effects of distinct moral perspectives within a society. For example, what kind of social ethic would emerge in a simulation of the Prisoner's Dilemma in which half of the population adopts a Kantian ethic model while the other half adopts a rational agent model? (Ruvinsky, 2008, p. 79).

Though she refers to the Prisoner's Dilemma, it seems like the simulation she envisions is an ABM with an interacting population and not a tournament in the style of Axelrod and Danielson. How one quantifies ethical theories using the parameters "rights, liberties, and duties" is not shown in any detail. I doubt that one can quantify ethical theories such as care ethics and virtue ethics based on these parameters alone. Nevertheless, Ruvinsky is noteworthy for being one of the few voices encouraging the computer simulation of ethical theories through agent-based modeling.

The first ever computer simulation of an ethical theory using ABM was done by Mascaro et al., as presented in their book, *Evolving Ethics: A New Science of Good and Evil* (Mascaro, Korb, Nicholson, & Woodberry, 2010). Their simulations were programmed in NetLogo, currently the most popular and accessible ABM software package. They first developed an evolving world where agents could move, eat, reproduce, and also pass on certain traits and behaviors to the next generation. Next, they used act utilitarianism to address the ethical status of controversial acts such as suicide, rape, and abortion in this evolving world. They chose act utilitarianism as the normative ethical theory because they think it is right and also because, according to them, it is "the only ethical system which *allows* us to measure the outcomes of computer simulations and judge them as better or worse" (Mascaro et al., 2010, p. 5). In other words, it is the only ethical theory that can be usefully quantified for computer simulations. All the other ethical theories, such as deontological ethics, virtue ethics, and even rule-utilitarianism "depend upon the exact semantics of the deontic principles or the virtues, respectively, and incorporating semantic understanding into artificial life simulation in any kind of sophisticated way requires a prior solution to the problem of natural language understanding" (Mascaro et al., 2010, p. 32). I strongly disagree with this point, and I hope to show in the next section that this is not the case. It may be that act utilitarianism is more *straightforward* to simulate than other ethical theories because what is minimally required is a quantification of the utility from every act, a single numerical variable. But other ethical theories can also be quantified without recourse to natural language understanding.

Mascaro et al. remind us that according to utilitarianism, what is good is what maximizes "the sum of expected utilities *across a population*" (Mascaro et al., 2010, p. 29). Though acknowledging that in real life, "what utilities themselves are is not exactly clear" (p. 27), they simplify things in their simulation by connecting utility to the variable, "health," which accounts for both physical and psychological health. An act that is committed is good if it produces greater utility (health) for the whole population than if the act was not committed. Fortunately, with computer simulations, this can easily be done and measured. A batch of simulations (perhaps with varying environmental conditions) can be run with act X turned on, then an equal number of simulations could be run with act

X turned off. The utility scores between the simulations can then be compared. In the case of Mascaro et al.'s work, simulations were run where suicide, rape, and abortion were present, and also where they were absent. These simulations spanned several generations of agents or thousands of digital years. Their findings reveal that suicide and abortion can in some extreme cases be ethical, namely when there is a scarce supply of food such as during a drought, while rape is always unethical because of the unavoidable health costs for the victim, both physical and psychological.

One could question aspects of their implementation. First of all, the values for negative and positive utilities must be determined by the programmer. "Those utilities are fixed, being selected to reflect the real world to some approximation" (Mascaro et al., 2010, p. 89). So for example, in the rape simulation, when a victim is raped, the victim derives a large negative utility of -70 health units (additionally, if there is offspring produced, the victim will have to invest anywhere from -590 to -110 health units as parental investment) and the rapist derives a small positive utility of 5 health units. If the rape is prevented according to a "rape prevention probability" variable (either 0.9, 0.75, or 0.5 depending on the experiment), then the targeted victim experiences a smaller negative utility of -10 health units while the rapist suffers a large negative health effect of -60 health units and negative utility of -15 health units, representing the rapist being punished (Mascaro et al., 2010, p. 186). Why a negative utility of -70 and not -100 or -150 for being a rape victim? Is it also reasonable to keep the negative utility constant in every case or is it better to introduce some fluctuations? Mascaro et al. acknowledge that in principle the value need not be fixed and can perhaps evolve over time (Mascaro et al., 2010, p. 96). However, this issue of setting appropriate values for utility presents a challenge for anyone who wants to simulate utilitarianism.

Second, whether it is "utility" (e.g. psychological trauma) or a more neutral "health effect" (e.g. giving birth) that is being referred to, they both involve "health units" and contribute to a single numerical variable called "health." This can be confusing because it is not clear whether utility is essentially different from health or the same as health. It seems that utility and health are distinct from the point of view of the simulator but not from the point of view of the virtual agent, who experiences both simply as health.

Despite these implementation problems, the greater value of the work of Mascaro et al. is its pioneering endeavor to simulate a specific ethical theory with ABM. Given the exploration of acts of suicide, rape, and abortion in multiple worlds and over many thousands of virtual years, it is obvious that one can do ethical experiments in computer simulations that one cannot do in real life. In fact, they call their project a "new science" and an "experimental ethics" that introduces a new methodology to the study of ethics. To the objection that their simulations might be too simple or naive, their answer is likewise simple: "go forth and simulate better!" (Mascaro et al., 2010, p. 236). Indeed, the controversial aspects of their work should be a spur to others to see how computer simulations of ethics could be better undertaken.

Prominent Ethical Theories

In a brief Internet article, Mike Loukides of O'Reilly Media speculates about "an AI that can compute ethics" and considers how the three major ethical theories of deontological ethics, utilitarianism, and virtue ethics might be considered "optimization problems" for a machine to solve (Loukides, 2017). He himself is skeptical about such a prospect but recognizes that these ethical theories indeed have features that are, in theory, computable. "It isn't surprising that computational ethics looks like an optimization problem. Whether you're human or an AI, ethics is about finding the good, deciding the best way to live your life" (Loukides, 2017).

I now turn to prominent ethical theories, namely deontological ethics, utilitarianism, feminist care ethics, and virtue ethics, and consider how they can also be simulated on the computer with ABM. I only provide pointers and suggestions for their simulation, not any technical implementation. But I hope that these will be enough to show that their simulations can be done.

Deontological Ethics

Kant's categorical imperative is as follows: "Act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant, 2002, p. 37). As Christine Korsgaard explains,

[Kant] suggests that the way to test whether you can will your maxim as a universal law is by performing a kind of thought experiment, namely, asking whether you could will your maxim to be a law of nature in a world of which you yourself were going to be a part... Kant's test may be regarded as a formalization of the familiar moral challenge: "What if everybody did that?" In order to answer this question, you are to imagine a world where everybody does indeed do that. (Kant, 2012, pp. xx-xxi)

Any maxim that passes this test counts as a duty, and a duty ought to be followed no matter what. It should not be influenced by any external factors, emotions, or unforeseen consequences.

Implementing this ethical theory in a computer simulation is challenging. One could at first suggest that it is easy to apply deontological ethics to the world already provided by Mascaro et al. To see whether suicide is a duty, simply have all the agents commit suicide in one simulation and observe what happens. But even in this trivial example, this would be a deontological ethics on the part of the experimenter who is not *part* of the virtual world itself. A unique aspect of deontological ethics is that it is an ethical theory that the *agent* doing the act needs to know and implement. The theory cannot be applied from *outside* as might be the case with utilitarianism or virtue ethics. In these other ethical theories, a person may act ethically (by maximizing net utilities or by performing virtuous acts) without consciously subscribing to utilitarianism or virtue ethics. In contrast, one cannot act ethically according to the categorical imperative without knowing it. It needs to be conscious and deliberate. This is also apparent in the

third formulation of the categorical imperative called the "formula of autonomy," which considers the will of the agent as the source or giver of universal law, i.e. the universal law cannot come from outside the agent (Kant, 2002, p. xviii).

Given this special condition, we need to add a more complex "cognitive architecture" to properly render deontological ethics. There are many cognitive architectures that have been developed for agents, one of the most well-known being the BDI (Belief, Desire, Intention) architecture (Rao & Georgeff, 1995).[2] However, without going into the details of any specific cognitive architecture, I suggest that the way to simulate deontological ethics is through a "simulation within a simulation." An agent must have the capacity to simulate another simulation in its head (a second-order simulation). Assuming that the agent subscribes to the categorical imperative, then before it performs a certain act (perhaps given to it as an option by the programmer or randomly generated), it must be able to imagine (simulate) a world where all other agents in its present world did the same act (for the sake of the example, let us assume that the agent is a powerful rational agent who knows everything about its present world).[3] This second-order simulation will be evaluated as either good or bad based on some standard. If it is good, then the agent will identify the said act as a *duty* and perform it.[4] If it is bad, then the act will not be performed.

How will the agent decide if the imagined world is good or bad? Should it be *better* in some way than the present world for it to be regarded as good? If so, in what way? If one considers the benefit to all the agents in the world then it would be similar to utilitarianism. My tentative answer is that the second-order simulation should be *sustainable* and *balanced*. Kant speaks of a universal maxim as a "universal law of nature," and nature usually tends to a kind of sustainable equilibrium. The imagined world should be *sustainable* in that the second-order simulation could continue for *n*-number of generations without any kind of catastrophe. A universal maxim of abortion obviously cannot be sustained because there would be no more agents by the next generation. On the other hand, a universal maxim of reciprocal helping could be sustained for an indefinite number of generations. How far ahead into the future the agent can look will depend on the programmer. Second, the imagined world should be *balanced.* This is more variable and could mean any number of things depending on what can be found in the first-order simulation. If agents possess wealth in the simulation, then *balanced* might equate to every agent having at least 1% of the total wealth in the world, with no single agent having more than 10%. A world where three agents ended up with 90% of all the wealth in the world would then count as *imbalanced*, something the rational agent would never consent to. Therefore, a universal maxim that leads to

[2] For a survey of cognitive architectures for agents, see Balke and Gilbert (2014).

[3] Because the second-order simulation depends on the state of the present virtual world (first-order simulation), this also gives some variability in what will count as a duty. What counts as a duty at one point could change over time as the conditions of the present virtual world change.

[4] There is also the complicating factor of how *often* the act should be performed as a duty. Every tick of the simulation? Once every cycle? For the sake of simplicity, we will ignore this issue here.

this kind of imaginary world would not count as a duty and will not be performed by the agent.

These suggestions can be refined further but I hope they show that there is a way of simulating deontological ethics which stays true to its special condition. It requires a "simulation within a simulation" or second-order simulation conducted by the deontological agent. Complex as it may sound, this arrangement is possible with current computing power and programming resources. It would be computationally taxing if we require that second-order simulations need to be conducted by agents for every single act. More efficiently, agents could just remember what they have identified as duties and reserve the second-order simulations for brand new acts. It could also cause significant slowness if we have many deontological agents doing second-order simulations at the same time. But if well executed, we could have a simulation with agents only doing acts that, according to their own reason, should be universal laws.

Utilitarianism

In their book, Mascaro et al. mention an interesting method for act utilitarianism:

Calculating the cumulative utilitarian effects of a specific act, e.g. a specific suicide, is straightforward enough. At the point in our simulation where the suicide occurs, we can fork the simulation, with one process containing the suicide and the other excluding it, and then compare the consequences. (Mascaro et al., 2010, p. 179)

They do not adopt this method because they claim they are more concerned with *kinds* of acts rather than individual acts, and also that this method would be impractical. However, this forking method is more faithful to act utilitarianism than the method that they adopt and should be developed by anyone wishing to construct a better simulation of act utilitarianism. The kind of act utilitarianism that Mascaro et al. have, which turns a particular action on or off for different simulations and then compares the results, is in fact not very far from rule utilitarianism. Rule utilitarianism in its simplest form states that an action is right if it follows a rule that leads to the greatest good. However, issues such as what it means for a specific action to *follow* a rule, what counts as a rule, and whether exceptions to the rule are permissible (especially those that maximize utility), have led to different versions of rule utilitarianism (Lyons, 1965). Without discounting the complexities and nuances involved, let us here consider a simplistic version of "strong rule utilitarianism" which holds that if a rule contributes to the greater good, it is always right to follow that rule and always wrong to break it.

Mascaro et al.'s simulations seem to already contain everything needed to explore this kind of strong rule utilitarianism. The effect of the rules, "one should not commit suicide," "one should not commit rape," and "one should not commit abortion" can already be observed in their simulations, namely, in simulations where those actions are turned off versus those simulations where they are turned on. However, the method of assessment for rule utilitarianism will be different from theirs since, as I understand it, rules are

more temporally independent than actions. Instead of looking at the net utility of a total population at a given time period (such as in a time of drought), one should look at the net utility of the total population over the *complete timeline*, i.e. the whole time span of the simulation. If the supertemporal, intergenerational net utility in the simulation where suicide is turned off is greater than the supertemporal, intergenerational net utility in the simulation where it is turned on, then the rule "one should not commit suicide" contributes to the greater good and the act of committing suicide is unethical in all cases, even in times of drought.

I will not dwell on this point because this strong rule utilitarianism is only a simplistic version and there are better and more nuanced versions of rule utilitarianism that cannot be discussed here. However, as mentioned in the previous section, I think where Mascaro et al.'s work can be improved is in the designation and assignment of utility. This will apply whether we use act utilitarianism or rule utilitarianism. Though they connect utility in their simulation with health, they recognize that there are many other sources of utility. For example, in the case of rape, the negative utilities are not only on account of the direct psychological and physical harm but also involve long-term trauma, negative utilities to the relatives and friends of the victim, etc. (Mascaro et al., 2010, p. 185). A more complex virtual world where long-term memory and emotions are included, as well as human social relationships, would allow for a better simulation of utilitarianism. In the last section, we discuss the PECS reference model as something that could work well with virtue ethics, but it could have benefits for utilitarianism as well.

Feminist Care Ethics

As a response to the two rational and *masculine* ethical theories mentioned above, Nel Noddings introduces a care ethics that she says is more *feminine* in its approach, an ethic that "has a proper regard for human affections, weaknesses, and anxieties" (Noddings, 2013, p. 25). According to her,

Caring involves stepping out of one's own personal frame of reference into the other's. When we care, we consider the other's point of view, his objective needs, and what he expects of us... Our reasons for acting, then, have to do both with the other's wants and desires and with the objective elements of his problematic situation. (Noddings, 2013, p. 24)

The concrete relationship between the "one-caring" and the "cared-for" is considered basic. In the caring relationship, the "one-caring" has an affection and regard for the "cared-for" that is not bound or dictated by rules. At the same time, the "cared-for" usually has an awareness of this affection and reciprocates in a proportional way. Noddings claims that care ethics is not a *theory* like utilitarianism or deontological ethics. It does not deal with the abstract and hypothetical but with concrete human relationships. It does not even claim universalizability for all human beings.

Ignoring the fact that Noddings would probably object to a computer simulation given its abstract and hypothetical nature, can care ethics be simulated? I suggest that a good place to start is to build on simulations that already simulate the begetting of offspring. In most of these simulations, offspring are practically the same as the parents, perhaps only with less health or a lower "age" value. There is no enduring link or relationship between the parents and children beyond the passing on of certain "genes," as in the case of evolutionary simulations. However, if we modify such simulations so that the child has serious weaknesses that the parent needs to address, and if we create a unique, enduring link between the two such that parent and child can identify themselves as the "one-caring" and the "cared-for" respectively, that could serve as a basic foundation for care.

For example, it would be the parent's responsibility to collect food in order to feed the child who would otherwise not survive. In order to simulate an awareness of the needs of the other, the parent must know the status of the child in terms of health. To simulate less of a linear rule-dictated behavior and more of an overarching "care" for the child, we can also imagine the parent behaving differently depending on the seriousness and urgency of the needs of the child. For example, the parent might collect food at a pace of 1 step/tick when it knows that her child has a good health of 50 units. But once the parent knows that her child's health has dropped dangerously low to 10 units, the parent might consider various options that she would not normally do: collect food at a pace of 3 steps/tick at the risk of exhaustion, explore a more dangerous part of the map for the sake of better food, etc., all with one goal in mind: to restore the health of her child. Interesting mass dynamics might be observed in crisis situations where many children start suffering from low health (e.g. famine, spread of disease among infants, etc.).

The reciprocity of care can also be simulated. When the parent reaches a certain age in the simulation, she could acquire handicaps and weaknesses. By this time, the child would be a strong adult and would be in a position to care for the aging parent. The roles of the "one-caring" and the "cared-for" will be reversed. This is in fact how it works in many traditional societies where children serve as the "insurance" of parents in their old age. After the caring relationship between parents and children is properly established and configured in the simulation, this can serve as the pattern for caring relationships between siblings, relatives and non-relatives.

What has been mentioned so far is the procedural part of caring. But if we want to be more faithful to what caring means to Noddings, then we would have to deal with the affections and emotions which add a deeper layer of complexity. The "one-caring" needs to be *engrossed* and perhaps even *sad* about the weaknesses and problems of the "cared-for." Meanwhile, the "cared-for" needs to know that the "one-caring" feels this way and that the caring acts are not perfunctory. There are many models proposed for the computer simulation of emotion, some of them based on actual psychological theories (Bourgais, Taillandier, Vercouter, & Adam, 2018). Whether any of these emotion models will suffice to simulate a caring relationship remains to be seen. But it seems that an emotional component is an essential requirement to genuinely simulate care ethics.

Virtue Ethics

Virtue ethics looks at the positive qualities (virtues) and negative qualities (vices) of persons that lead to habitual actions and behavior. It traces its origins to Aristotle's *Nicomachean Ethics* and experienced a revival in the 20th century through the work of Alasdair MacIntyre (MacIntyre, 1981). Perhaps the greatest advantage of virtue ethics is that it is essentially "agent-based." It looks primarily at the person before looking at the person's actions.

Though no author explicitly refers to a computer simulation of virtue ethics, Coelho et al. imply it. They talk about designing an intelligent agent with character and virtues, where "agency and character [virtue is a character trait] merge together and are responsible for all the behaviours generated" (Coelho, da Rocha Costa, & Trigo, 2014, p. 22). In order to do this, they designed an agent that operates a stochastic game in its moral decision-making process, namely, "a partially observable Markov decision process or POMDP" (Coelho et al., 2014, p. 24). Without going into the details of their work, I think the idea of employing stochastics and probability in a virtue ethics simulation will be useful. When we say that a person is just or possesses the virtue of justice, we do not mean that the person does just actions *all* the time, though that person might be expected to do just actions *most* of the time. In simulation terms, an agent might have a justice value of 70%, which means we can expect the agent to perform just acts in about 70% of cases where such acts could apply.

Coelho et al. admit that "it is not easy to design an agent to be gracious, merciful or respectful" (Coelho et al., 2014, p. 22). I add that the success of simulating virtuous agents will depend on their capacities in the virtual world. It is perhaps not so easy to simulate virtues and vices when agents can only eat, move, and reproduce. However, in some simulations such as the classic *Sugarscape* of Joshua Epstein and Robert Axtell, agents can also possess wealth, trade with each other, and come into conflict (Epstein & Axtell, 1996). Such a world would be more conducive for virtues and vices. For example, let us say that when two agents conduct a trade of different resources between them, there is a small probability for an opportunity to "cheat" or "steal" to arise, i.e. an opportunity to take resources illicitly from the other agent. When this opening comes up, a just agent might have a 70% chance of declining this opportunity to cheat, while an unjust agent might only have a 10% chance of "resisting the temptation." We can imagine that as the unjust agent experiences the thrill and reward of cheating, its vice of injustice strengthens, decreasing its justice level to 8%. On the other hand, the just agent also increases its justice when it is able to decline the cheating opportunity, perhaps to 72%.

An additional mechanism is the probability of being found out and punished (in the case of the cheater) or praised (in the case of the just agent). Being punished might teach the cheater a lesson and discourage it from cheating again in the future (increasing its justice level to 12%). Being praised might encourage the just agent to become even more honest

(raising its justice level to 75%). All the dynamics mentioned can easily be simulated since they only involve simple probabilities.

I can also offer a more abstract approach for the simulation of virtue ethics. Let us assume that agents tend to find themselves in certain situations represented by a mathematical function. Each situation can be addressed with an abstract virtue A. An agent will perform gradient descent *n*-number of steps on the situation-function, with *n* determined by the level of their virtue A. The closer the agent comes to a local minimum of the situation-function, the more it can be said that the agent has addressed the situation "virtuously." If it reaches the global minimum, then the agent can be said to have acted in the most perfect way possible given the situation. Conversely, we could talk about gradient ascent and maxima for vices. Reaching either a minimum or maximum of the function can be bound with certain rewards or punishments for the agent. Situation-functions can resemble each other in various ways or can mutate. Some situation-functions might be "tougher" than others, requiring a greater degree of virtue to "solve," i.e. to find a minimum.

The disadvantage of such an abstract approach is that it is not apparent what situation, virtue, or vice from real life is being represented, unlike in the previous cheating example. On the other hand, the advantage is that a situation-function could represent practically any situation in human life, such as "buying a car (requiring prudence)," "confronting a bully (requiring courage)," etc.

Whether we choose a more concrete or abstract approach, if we want to accurately portray virtue ethics then we need to account for emotions just like in care ethics. In traditional virtue ethics, a virtue is considered a virtue if it allows the higher rational part of the agent to control and direct the lower instinctive and emotional part which can often *resist* this control (Aquinas, 2010, pp. 18-19). Emotions need to be simulated to more faithfully depict virtue ethics. Furthermore, in the cheating example above, I also mentioned the possibility of being punished or praised by others which is a social dimension to virtue. A virtue (or vice) can flourish if society at large praises and rewards those who have it. So a system of social reputation needs to be introduced. These emotional and social components will be reiterated in the next and last section when we introduce the PECS reference model for simulation.

To conclude, it is not true, as Mascaro et al. have stated, that natural language in AI is required to simulate other ethical theories besides act utilitarianism. In fact, rule utilitarianism could already be applied to their simulation simply by using a different kind of perspective for assessment; deontological ethics can be simulated using the technique of a "simulation within a simulation" or second-order simulation; the basic mechanism for care ethics can be built on top of simulations that involve parents and children; and virtue ethics can be simulated simply with probabilities or more abstractly with mathematical functions and a gradient descent algorithm. The recurring challenge is the inclusion of emotions, which is needed in care ethics and virtue ethics and to a

certain degree also in utilitarianism. However, there are already emotion models for simulation that can be explored (Bourgais et al., 2018). So far, we have not encountered any insurmountable barriers to the computer simulation of ethical theories. In the next section, I point out the compatibility of virtue ethics with a simulation model that has already been proposed, called PECS. This is a precursory step to attempting the technical implementation of a computer simulation of virtue ethics.

Virtue Ethics and the PECS Reference Model for Simulation

The PECS Reference Model was introduced by Bernd Schmidt and Christoph Urban for the simulation of human behavior in a social environment. It stands for **P**hysical Conditions, **E**motional State, **C**ognitive Capabilities, and **S**ocial Status (Schmidt, 2000, p. 1). It was proposed as an alternative to the popular BDI (Belief, Desire, Intention) model. BDI is a good framework for the rational decision-making process but does not account for emotions and common social behaviors such as communication and learning. Schmidt and Urban claim that to more accurately model human behavior, these components should be included. However, they only provide concepts and guidelines for such a model and not a technical implementation. Based on the literature, PECS has definitely not replaced BDI and it has not received any enduring support from the ABM community. Nevertheless, I think it is naturally compatible with virtue ethics and should be used as the foundation for a computer simulation of virtue ethics.

As was mentioned above, in traditional virtue ethics, a virtue is described as the higher rational part controlling and guiding the lower instinctive and emotional parts of the soul which may *resist* the higher part. This dichotomy corresponds to the two kinds of behavior in PECS: "deliberative behavior" and "reactive behavior," with the latter including "drive-controlled behavior" and "emotionally controlled behavior" (Schmidt, 2000, p. 1). Presumably, "deliberative behavior" comes from the cognitive component of the agent, whereas "reactive behavior" comes from the physical and emotional components. According to Schmidt and Urban, there must be a constant interaction between these components.

In addition, virtues do not arise out of thin air. They are learned from others. A child usually learns them first from parents and elders, and then later on from peers and famous "exemplars."[5] As Aristotle says, "it is not unimportant how we are habituated from our early days; indeed it makes a huge difference – or rather all the difference" (Aristotle, 2000, p. 24). So virtues (and vices) should be able to be "passed on" from one agent to another. As mentioned in the preceding section, virtues (or even vices) can also thrive if they are praised by a particular community. If courage is a highly praised virtue in society, then an agent who is especially keen on social praise will have a greater inclination to be courageous. In short, agents are encouraged towards certain behaviors

[5] For a discussion of the importance of moral exemplars, see *Exemplarist Moral Theory* (Zagzebski, 2017).

because of social influence. All these elements of virtue ethics would fall under the social component of the PECS model.

To provide an example, temperance is the virtue of moderation in matters of physical desire such as food, drink, and sex (Aquinas, 2005, p. 119). We can imagine an agent with a weakness for food who eats more food than everyone else, perhaps 3 units of food instead of the normal 2. This physical drive is genetic, i.e. it is programmed into the agent. Let us say that this is compounded by an emotional response: when the agent is unable to regularly eat 3 units of food, the agent gets "angry" with an irrational proneness (probability) to attack one of its neighbors. This would lead to serious punishment and negative consequences for the agent. However, if the agent has the virtue of temperance, the virtue would have the probability of curtailing these weaknesses either by addressing the physical weakness (it makes the agent eat 2 instead of 3 units of food) or the emotional weakness (it prevents the anger from arising when the agent is hungry). Though on the outside the agent seems to act like everyone else, the agent is more virtuous because it actively practices a virtue that regulates its imbalance. If we imagine a society that praises the virtue of temperance, then this could encourage the agent to increase its level of temperance even more depending on how much the agent is receptive to social influence. This is a relatively simple example, but it touches on all the components of the PECS model.

To conclude, there already exists a reference model for social simulation that is conducive for simulating virtue ethics. As its authors state, "it is a fundamental conviction of the PECS research program that an understanding of human behaviour can be achieved only if all 4 aspects and their interaction are taken into account" (Schmidt, 2000, p. 20). Virtue ethics, in its traditional form, also requires all four aspects and their interaction to be taken into account. What remains is to attempt a computer simulation of virtue ethics using the PECS model as its foundation.

Conclusion

It is possible to do computer simulations of ethical theories with ABM. Though there are challenges involved, there are no insurmountable obstacles to such an endeavor. I hope that this paper provides a starting platform for philosophers to use computer simulations just as many social scientists are currently doing. Computer simulations provide the opportunity to conduct virtual experiments impossible in real life due to practical or ethical considerations. They can be considered "thought experiments" with greater degrees of detail and sophistication than what can ever be done with words. A computer simulation can provoke a deeper understanding and appreciation of an ethical theory as its unique aspects are hammered out in the programming process, or as its consequences are observed in visual simulation. If the simulation is expertly constructed, it might even contribute to defending an ethical theory as being universally valid for human experience, real and imagined.

References

[1] Aquinas, T. (2005). The Cardinal Virtues: Prudence, Justice, Fortitude, and Temperance (R. J. Regan, Trans.). Indianapolis: Hackett.

[2] Aquinas, T. (2010). Disputed Questions on Virtue (J. Hause & C. E. Murphy, Trans.). Indianapolis: Hackett.

[3] Aristotle. (2000). Nicomachean Ethics (R. Crisp, Trans.). Cambridge: Cambridge University Press.

[4] Axelrod, R. (1984). The Evolution of Cooperation. New York: Basic Books.

[5] Axelrod, R. (1997). The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration. Princeton, NJ: Princeton University Press.

[6] Balke, T., & Gilbert, N. (2014). How Do Agents Make Decisions? A Survey. Journal of Artificial Societies and Social Simulation, 17(4)(13). doi:10.18564/jasss.2687

[7] Bourgais, M., Taillandier, P., Vercouter, L., & Adam, C. (2018). Emotion Modeling in Social Simulation: A Survey. Journal of Artificial Societies and Social Simulation, 21(2)(5). doi:10.18564/jasss.3681

[8] Coelho, H., da Rocha Costa, A. C., & Trigo, P. (2014). On Agent Interactions Governed by Morality. In D. F. Adamatti, G. P. Dimuro, & H. Coelho (Eds.), Interdisciplinary Applications of Agent-Based Social Simulation and Modeling (pp. 20-35). Hershey, PA: Information Science Reference.

[9] Danielson, P. (1992). Artificial Morality: Virtuous Robots for Virtual Games. London: Routledge.

[10] Doloswala, K. N. (2014). Eroding trust - An agent based model to explore how trust flows. Australasian Marketing Journal, 22, 51-53.

[11] Epstein, J. M., & Axtell, R. (1996). Growing Artificial Societies: Social Science from the Bottom Up. Washington D.C.: Brookings Institution Press.

[12] Gilbert, N., & Troitzsch, K. G. (2005). Simulation for the Social Scientist (Second ed.). Berkshire: Open University Press.

[13] Kant, I. (2002). Groundwork for the Metaphysics of Morals (A. W. Wood, Trans.). New Haven: Yale University Press.

[14] Kant, I. (2012). Groundwork of the Metaphysics of Morals (M. Gregor & J. Timmerman, Trans. M. Gregor & J. Timmerman Eds. Revised Edition ed.). New York: Cambridge University Press.

[15] Kim, W.-S. (2009). Effects of a Trust Mechanism on Complex Adaptive Supply Networks: An Agent-Based Social Simulation Study. Journal of Artificial Societies and Social Simulation, 12(3)(4).

[16] Lim, H. C., Stocker, R., & Larkin, H. (2008). Ethical Trust and Social Moral Norms Simulation: A Bio-Inspired Agent-Based Modelling Approach. Paper presented at the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Piscataway, N.J.

[17] Loukides, M. (2017). On computational ethics: is it possible to imagine an AI that can compute ethics? Retrieved from https://www.oreilly.com/ideas/on-computational-ethics

[18] Lyons, D. (1965). Forms and Limits of Utilitarianism. New York: Oxford University Press.

[19] MacIntyre, A. (1981). After Virtue: A Study in Moral Theory. London: Duckworth.

[20] Mascaro, S., Korb, K. B., Nicholson, A. E., & Woodberry, O. (2010). Evolving Ethics: The New Science of Good and Evil. Exeter: Imprint Academic.

[21] Noddings, N. (2013). Caring: A Relational Approach to Ethics and Moral Education (Second ed.). Berkeley: University of California Press.

[22] Rao, A. S., & Georgeff, M. P. (1995). BDI-Agents: From Theory to Practice. Paper presented at the Proceedings of the First International Conference on Multiagent Systems, San Francisco.

[23] Ruvinsky, A. I. (2008). Computational Ethics. In M. Quigley (Ed.), Encyclopedia of Information Ethics and Security (pp. 76-82). Hershey, PA: Information Science Reference.

[24] Schelling, T. (1978). Micromotives and Macrobehavior. New York: W. W. Norton.

[25] Schmidt, B. (2000). The Modelling of Human Behaviour. Ghent, Belgium: SCS-Europe BVBA.

[26] Tykhonov, D., Jonker, C., Meijer, S., & Verwaart, T. (2008). Agent-Based Simulation of the Trust and Tracing Game for Supply Chains and Networks. Journal of Artificial Societies and Social Simulation, 11(3)(1).

[27] Wilensky, U., & Rand, W. (2015). An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo. Cambridge, Massachusetts: Massachusetts Institute of Technology.

[28] Zagzebski, L. T. (2017). Exemplarist Moral Theory. New York: Oxford University Press.