Artificial moral agents: creative, autonomous and social. An approach based on evolutionary computation

Abstract

In this paper we propose a model of artificial normative agency that accommodates some crucial social competencies that we expect from artificial moral agents. The artificial moral agent (AMA) discussed here is based on two components: (i) a version of virtue ethics (VE); and (ii) an implementation based on evolutionary computation (EC), more concretely genetic algorithms. The reasons to choose VE and EC are related to two elements that are, we argue, central to any approach to artificial morality: autonomy and creativity. The greater the autonomy an artificial agent has, the more it needs moral standards. In the virtue ethics model each agent builds its own character in time. In this paper we show how both VE and EC are more adequate to a "social approach" to AMA when compared to the standard approaches based on deontological or consequentialist models implemented through standard computational tools.

The project of an autonomous and creative Artificial Moral Agent (AMA) thus implemented is the GAMA=genetic-inspired autonomous moral agent. First, unlike the majority of other implementations of machine ethics, our model is agent-centered; it emphasizes the developmental and behavioral aspects of the ethical agent. Second, in our model the moral agent does not follow rules (deontological model) or calculate the best outcome (consequentialism) of an action, although we incorporate rules and outcomes as starting points of our computational model (as initial population of the genetic algorithms). Third, our computational model is less conventional, or at least it does not fall within the Turing tradition in computation. Genetic algorithms are excellent searching tools that can avoid local minima and generate solutions based on previous results. In this respect in our GAMA model, the VE approach to ethics is better implemented by EC. As a philosophical aspect of our project we discuss the hybrid aspect of our implementation , *viz.* (Allen, Smit, & Wallach, 2005). Finally, and this is our main focus in this paper, we show that as social agents, the GAMA agents integrate. Finally, when appraised against the more widespread choice: the rule-based or calculation-based ethics implemented in more Turing-like architecture, the GAMA is more societal and .

The human-machine demarcation

As technology advances, some activities, once deemed as "human-only", are performed by machines or, more generally, artificial agents (we include here robots, algorithms, machines, etc.). As they accomplish more and more complex tasks: driving cars or school buses, flying airplanes or drones, fighting on battlefields, recognizing human emotions and expressions, trading high-frequency stocks, etc., something more than complex computational capabilities are needed for artificial agents: some minimal moral capacities to decide about moral matters. But can artificial agents distinguish right from wrong?

Artificial non-moral agents are already a reality of our everyday life; they are both needed and inevitable. Artificial systems outperform us in many areas: computers beat the majority of chess masters; in some cases, expert systems put a more accurate diagnosis than human doctors; for large datasets, computers are able to see patterns not visible even to experts.

Are artificial moral agents possible? Some philosophers of cognitive science as well as computer scientists are somehow optimistic and argue that sooner or later the concept of ethical agents is going to expand to include the artificial moral agents (AMAs). Being a moral agent is a matter of degree, depending on how many human-like faculties and actions the artificial agent is able to implement. A cautionary attitude would be to admit that some human faculties cannot be implemented in artificial structures. Let us suppose for the sake of the argument that one can identify a set of faculties F and a set of actions A, such that there is a sharp demarcation between what an artificial system can have and do (when compared to humans):

H-1 *DEMARCATION PRINCIPLE. There is a set of faculties $F=\{F_1 \ldots F_n \ldots\}$ that belongs to humans only and not to artificial systems (actual or possible). Consequently, there is a set of actions $A=\{A1 \ldots An \ldots\}$ that only human agents can perform, and not artificial agents (actual or possible).*

---

3 We follow the existing terminology (Allen, Smit, & Wallach, 2005; Allen, Varner, & Zinser, 2000; Allen & Wallach, 2009)

If H-1 is true, there are pure human cognitive faculties and pure human actions. The existence of such a demarcation principle is a topic that deserves a close philosophical and scientific scrutiny. The history of computer science and especially the development of AI systems, can be read as an attempt to reduce the sets *F* and *A*. The faculty of using natural language, free will, autonomy of agency, scientific and artistic creativity, religious experience, and, last but not least, being social and moral *with and to humans* are among the best candidates for the set F of human-only faculties.

The optimistic attitude, as opposed to the more cautionary one, is to assume that *in principle* sets *F* and *A* are empty, although we are far from implementing all faculties *F* and actions *A*: technological progress and advancements in cognitive science will provide a solid basis to reject any demarcation between human and the machine faculties and actions. As a more foundational question, we ask: what accelerates the process of implementing more faculties in F and more actions in A? The answer we look for here involves two capacities which are usually deemed human-only: intelligence creation and autonomy.

First, intelligence creation, or the capacity to design and implement intelligent systems is a member of F, not yet implemented in artificial systems. We are able to program and create new machines with more faculties F_i… etc., able to perform more actions A_j. Chalmers calls this the "proportionality thesis": an increase in human intelligence will produce an increase in intelligence creation. 21 (Chalmers, 2010).

Are AI systems able to create intelligence? can they create other AI systems better than we do? Imagine that some AI systems would sometimes surpass our faculty to create AI systems and that it starts to create new AI systems. This would be a tipping point, or a "singularity moment", when AI systems will be more reliable and more competent in creating new artificial systems than the average humans (Chalmers, 2010; Kurzweil, 2006).

Second, there is another essential faculty among F that connects the faculties in F with actions in A: autonomy. We have a higher degree of autonomy when it comes to our human-only faculties and actions. Are machines actually autonomous in performing different actions, even if they have the necessary faculties to do so?

Moral agents: autonomous, creative, social

Central to our argument is to admit that one member of F directly affected by computational creativity and autonomy is moral reasoning. We are moral reasoners and moral agents. Is a machine able to compute and choose the best *moral* action? The question whether ethics can programmed into a robot or whether mature, human, ethical competence can be emulated or simulated by what is now often termed an "artificial moral agent" (AMA) has been around for a long time, at least since Isaac Asimov first promulgated "three laws of robotics" (Asimov, 1942). Although the first attempts at design and coding date back to early 1990s (Danielson, 1992), interest in the topic and work on design and coding problems have intensified in recent years, driven partly by the rapid push toward autonomy in weapons systems and automatic pilots or drivers. The pioneering work of Ronald Arkin, who is passionately committed to the goal of building autonomous robots that are more moral than human combatants has significantly advanced the field (R. C. Arkin, 1998; R. Arkin, 2009). As Wallach and Allen have argued, we are at a point where it is now a necessity that we implement ethics modules in wide range of autonomous systems (Allen & Wallach, 2009). For Allen and collaborators, there is an urgent requirement to enlarge moral reasoning to machines, given the demand and supply of such machines, and their political and legal consequences. From this one can ask another question: are moral autonomous robots supposed to display social skills when interacting with humans?

Skeptics may ask more foundational questions: Are moral and ethical judgments, based on values, computable? And by extension, are social skills computable in principle? Can machines do ethics and debate the optimal outcome in an ethical context, let us say a moral dilemma, in the same way in which they play chess or pilot an airplane? What degree of moral autonomy can we confer to a machine?

A general answer to all these questions is hard to give. From a philosophical point of view, an inquiry into machine ethics may clarify some aspects of the human-machine interface and some aspects which are common to humans and machines: what is a moral action? what is a moral judgment? is ethics based on rule-following reasoning? and last but not least, what is moral autonomy in a social context?

When the social aspects of AMA is abstracted away, trivial answers are probably easy to fathom: if ethics is simply a procedure of rule-following with no or little autonomy and creativity, then machine ethics looks like much easier to implement. A set of moral rules are programmed into a machine and the machine follow them blindly. Moral principles are simply conditions imposed on the structure of the algorithm and social interaction with humans is less important. The moral robot is always ready to act the way it was programmed and learning as well as developing a personal habit are out of question.
Deontological and consequentialist models of ethical behavior may or may not fare well in respect of the social aspects of AMA. Is an ethical action all about following a set of moral principles or rules? Are we ethical when we follow a set of rules? Are social values reducible to rules?

Social interaction, as described in social psychology, is hard to code in the algorithms of an AMA. It is the aim of this paper to suggest that Virtue Ethics applied to AMA may mesh better with the social aspect of moral agents. First, one assumption adopted here is that both intelligence creation and autonomy are central in any discussion on the demarcation between us—as moral reasoners and agents—and machines. Another assumption is that the decision making in machine ethics is *not* co-extensive with rule-following. In other words, we are doing *something more* than following rules when we reason and act morally. The third assumption is that a machine ethics centered on the moral agent and its character, is preferable to an ethics based on action and moral reasons.
Philosophers tried to find an alternative to rule-following ethics. Probably the most promising ethical theory that has challenged the rule-following framework is virtue ethics. The individual character of the virtue ethics model is related to other individuals. The social dimension of ethics is different, if not better in the case of virtue ethics.

Explicit moral agents

Some classifications are in order here: we rely on recent work in the philosophy (both epistemology and ethics) of machine ethics: (Allen, Varner, & Zinser, 2000; Allen & Wallach, 2009; Anderson & Anderson, 2011). J. Moor, in a recent introduction to machine ethics, ("The Nature, Importance, and Difficulty of Machine Ethics") discusses several varieties of artificial moral agents (Anderson & Anderson, 2011). Two of them are: the "implicit moral agents", programmed to behave ethically by following ethical principles, but constrained by their designers; and the "explicit ethical agents" which compute the best ethical action, but in an altogether different manner: the explicit artificial agent is able to assess the output based on some ethical theory and take the best action. The explicit moral agent has a form of autonomy. For explicit moral agents, the challenge to programming are much higher and J. Moor, together with J. Gips, would agree that developing an ethical robot is the Grand Challenge of computing (Anderson & Anderson, 2011).

For or against machine ethics

Autonomy and creativity of artificial agents can be seriously questioned. Are computers really creators and discoverers, or are they mere tools used by humans? Whereas computerized works of art existed for a while and computers are used frequently to assist the process of scientific discovery, the artificial moral decision is a more problematic area.
The technical terms that can be used for such an endeavor is machine ethics or artificial morality (Allen et al., 2005, 2000; Anderson & Anderson, 2011). Is the concept of machine ethics at least coherent? For many, probably a majority, machines are not capable to make moral decisions. Even if we grant them this ability, some others would argue that this is not desirable for social and moral reasons.

Allen and Wallach write that AMA implementation "highlights the need of a for a richer understanding of human morality" (Allen & Wallach, 2009, p. 216). Although synergy between human morality and AMA is desirable, it is not methodologically the only path possible. When we apply genetic algorithms to scientific discovery for example we do not simulate the process of human scientific discovery, but we aim to the similar results. The project is closer in spirit to the framework of artificial life than to the more restrictive domain of artificial emulation of cognitive processes.

Directly related to the skeptical position sketched above, one can ask this questions:

Q-1 *COMPUTABILITY: Is the moral decision making process computable? In particular, is the decision modelled by virtue ethics computable?*

We believe that in respect of artificial intelligence, the reduction of moral agency to humans is transient at best, if not fundamentally flawed. What reasons do we have to assume that moral agents need to be human? We focus on the idea of autonomy and start from the idea that there are, at least in principle, Artificial Moral Agents (AMA) which act autonomously.[3] Autonomy and morality are related: as an agent becomes more autonomous, it needs more moral standards.[4] But this connection between autonomy and morality already suggests that a rule-following ethics may not be the best choice. AMA are built to assist, or ultimately replace, human moral agents in a variety of situations: self-driven cars, autonomous weapons, drones, rescue missions, etc. If the emphasis is put on the autonomy of the AMA, it is weird to see how a proposal to stick with rules and principles can be successful.

Our project aims to an individual based ethics, more precisely a character based ethics, which is in fact the virtue ethics model adopted.

Two assumptions and the genetic-based autonomous moral agent (GAMA)

We challenge here the idea that moral decision making and moral agency belong to the sets F and A. We start with the assumption that we can have *in principle* AI systems that can make moral decisions in a more or less autonomous way and that there is normative and moral agency. This, at its turn is related to social competencies of AMAs.

The question is whether we can grant *enough* autonomy to artificial agents. Can we imagine a moral agent that has enough autonomy, compared to a human agent? A soldier in a field needs to make immediate decisions without asking the superiors.

We are interested in the foundational issues related to morality of non-human agents, in adopting the best ethical model for AMA *and* the best computational technique to implement it. We adopt a pluralistic view about moral agents. We do not assume that there is one, human-like moral decision procedure: in various contexts, moral decision making process can be more or less close to the human-like moral process. AMA can mimic humans or, depending on the context and content of the moral problem, they can deviate from the human decision making process or even avoid it intentionally. As Allen et al. suggest, a "moral Turing test" is needed to decide on a case-by-case basis whether AMA is performing better than (any) human moral agent.[5] Machines can mimic the way we decide morally or may not mimic us, but what we are interested here is a better result than what the one of the average human agent. Or probably we can reformulate Searle's Chinese Room argument against AI into an argument against a genuine understanding of what AMA decides (Gips, 1995).

Although these questions are enticing, at least philosophically, one can imagine a different approach. The question about the ability of the AMA to make ethical decisions has to be framed into an ethical theory. A better question would be: how do we create AMAs that given a certain ethical model, can surpass the average (human) moral competency within that model? As it is suggested in the literature, we have to relate our approach to AMA to a given ethical model, either in a top-down model, in a bottom up model or a hybrid model.

First, we answer Q-1 by adopting this working hypothesis:

H-2 *COMPUTATIONAL ETHICS: Moral decision making procedures can be modeled by a computational process.*

We assume that the moral decision making resembles a natural process that can be modeled and simulated by a computational technique. Then, there are two major novel hypotheses of our project: the type of computation used and the type of ethical model.

AMA implementation through evolutionary computation

It is first of all the implementation of an AMA which raises some fundamental questions about computation and representation. One issue is whether we should implement AMA as a rule-following algorithm, or we should follow

---

[3] We follow the existing terminology (Allen, Smit, & Wallach, 2005; Allen, Varner, & Zinser, 2000; Allen & Wallach, 2009)

[3] We follow the existing terminology (Allen, Smit, & Wallach, 2005; Allen, Varner, & Zinser, 2000; Allen & Wallach, 2009)

[4] The idea of freedom and autonomy are independent. See the quote in (Allen & Wallach, 2009, p. 23)

[5] Different moral dilemmas can be used in such a Moral Turing test: the trolley problem for example is the first to come to mind (Allen & Wallach, 2009, 1)

a different path in which rule-following does not play a central role. We believe that this second path is the evolutionary computation implemented by genetic algorithms.

H-3 *Evolutionary computation in ethics: The best technique used in ethical programming is evolutionary computation.*

One can think of the AMA implementation problem as a pragmatic problem and more or less as an empirical question in computational science. We will build a case for H-3 based on some similarities between evolutionary computation and the ethics needed for AMA. This is related to the second component of the project, its ethical model.

Genetic algorithms start from a given number of initial individuals randomly distributed in a given space, called the initial population. The genetic algorithm transforms individuals, each with an associated value of fitness, into a new generation by using the principles of survival-of-the-fittest, reproduction of the fittest and sexual recombination and mutation. Similar to Wright's landscape, the genetic algorithm finds "the most suitable" or the "best so far" solution to the problem by breeding individuals over a number of generations.

The procedure can be stopped by a termination condition: when the sought-for level of optimality is reached or when all the solutions converge to one candidate. The fitness function estimates the fitness to breeding of individuals in accordance with the principle of survival and reproduction of the fittest:

- Better individuals are more likely to be selected than inferior individuals.
- Reselection is allowed.
- Selection is stochastic.

The genetic algorithm ends with a *termination condition*: the algorithm can find a successful individual or complete a maximum number of steps. The success predicate depends on a choice made by a human and can be deemed as a pragmatic criterion. The winner is designated at the "best-so-far" individual as the result of the run.

AMA and its ethics

We start with the remark that the best ethical approach to AMA may or may not fit the best ethical approach to human agents. The standard proposal we have heard of is to adopt a combination of deontology and consequentialism ethics (Allen et al., 2000). As we argue here, virtue ethics for AMA is a promising alternative to the consequentialism or deontology.

H-4 *VIRTUAL VIRTUE: The approach to AMA based on a version of virtue ethics is better than other approaches based solely on consequentialism or deontology.*

We argue for H-4 by comparing and contrasting other approaches to our new proposal. They form a "package" and our claim is that virtue ethics and genetic algorithms are conceptually and foundationally appropriate to implement an AMA. These two issues are strongly related to the very general problem of autonomy and creativity of AI systems in respect of ethical decisions.

Why virtue ethics?

Hypotheses H-3 and H-4 complement the existing approaches in the literature on ethical programming. Originally, the problem of ethical computing has been integrated into the more general problem of epistemology of artificial systems and especially the rationality associated to them (Danielson, 1992; Ford, Glymour, & Hayes, 1995). The idea was that ethical machines *learn*, therefore they acquire information and make decisions based on such information, either by imitating humans or by following a rational process of choice between various alternatives. For the majority of authors, ethical programming is fundamentally related to the learning algorithms and to the acquisition of data about human behavior and therefore is ruled based and action based. For example, P. Danielson's earlier approach to build a better moral robot entails building more rational machines that perform better in a rational choice game (Danielson, 1992). Danielson's original approach contained an evolutionary component, mostly based

---

6 Genetic change is but not limited to: mutation, sexual recombination, gene flow, duplication, but it was not well understood till the mid-1970s. In the 1980s, the amount of data available gathered in some sciences (geophysics, meteorology, astronomy, particle physics, neuroscience) grew exponentially together with the computer power. New techniques of data mining and pattern discovery were needed.

7 I take here algorithms as abstract, mathematical objects, whereas programs as their concrete instantiation on a machine. A sensitive difference is between genetic algorithms, genetic programming and genetic strategies. See (Jong, 2006).

8 The names of these evolutionary models are not very informative.

on the social interactions of robots. When these machines are confronted with paradigmatic cases, they perform a utilitarian calculus or follow a set of predetermined rules similar to the deontological model of ethical decision.
Our project is not essentially based on consequentialism or on deontology (the leading suggestion in Danielson and Anderson). Some questions surrounding consequentialism and deontology that are relevant to AMA are: do we act morally when we simply follow rules? What is morality when there are no more rules to follow or lessons to learn? Do we need to develop morality of the individual or morality is associated to each act alone? What do we do when different moral rules are inconsistent? In philosophy, these are known difficulties for both consequentialism and deontology.
The third model is virtue ethics, which is not essentially an action-based ethics, but an ethics based on individual character. In contrast to deontology or consequentialism, individuals are no more performers-of-actions but they bear moral traits. What matters is the consistency of conduct, where conduct is understood as a feature of a set of actions of the same individual. Many philosophers would praise virtue theory for its practicality, when compared to deontology or even utilitarianism. Morality is about acquiring a number remarkable virtues, not necessary a unity of virtues, but a plurality of "moral traits".[9]
The most basic idea of virtue ethics is that moral excellence is more fundamental than moral obligation and rightness of action. Virtue ethics is centered on the individual and on his or her life achievement of a moral character. For A. Macintyre, virtues are grounded in practice. A practice is defined as "any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and partially definitive of, that form of activity, with the result that human powers to achieve excellence, and human conceptions of the ends and goods involved, are systematically extended." (MacIntyre, 1984, p. 187) Exercising a virtue is partly constitutive of excellent practice. The problem this type of virtue ethics faces is: how are we to decide which practice is worth pursuing and which practice to abandon?
There are several answers available to this question. One answer to this problem is to emphasize the contextual character of virtues as opposed to more universal and unitary role of virtues in Aristotelian ethics.
A virtuous agent is one who "has and exercises virtues"(Hursthouse, 1999). R. Hursthouse adds that right actions are actions a virtuous agent would perform in the relevant circumstances. Take for example the morality of an abortion. For Hursthouse the variation which makes an action right or wrong depends on the differing motives and thinking of the woman who has the abortion in that specific circumstance, rather than on the consequences of this act in all circumstances.
The other answer is that right actions call for moral wisdom, which is acquired through experience, and is not typically found in un-experienced agents.

The elements of an artificial virtue ethics

How do we relate virtue ethics to artificial agents? Our hypothesis H-4 has not been scrutinized enough philosophically. Computer scientists and cognitive scientists remarked the interesting link between virtue ethics and the process of learning. Also given its natural explanation of the process of moral development, virtue ethics may well be the most naturalized form of ethics. Hence its ability of being implemented in artificial agents. Recently, Gips (Anderson, Anderson, & Gips, 2011) suggested that a virtue ethics model adapted to AMA is promising and linked it to connectionism: "both seem to emphasize the immediate, the perceptual, the non-symbolic. Both emphasize development by training, rather than by the teaching of abstract theory."[10] In the 1990s, some philosophers (Paul and Patricia Churchland among others) suggested that a connectionist neural networks (CNN) model is suitable to approach moral cognition (Churchland, 1995). It is interesting to see how some philosophers who want to naturalize morality and ethics have adopted the virtue ethics model, rather than the more popular consequentialism or deontology (Casebeer, 2005).
Another suggestion that virtue ethics may be more suitable to AMA comes from C. Allen and his collaborators who proposed a hybrid approach to AMA based on virtue ethics (Allen et al., 2005; Allen & Wallach, 2009, Chapter 5).

---

[9] We do not need to adopt here for the ethics of AMA the controversial thesis that all virtues form a unit.

[10] This is a revised version of (Gips, 1995).

The main distinction they drew is between the top-down approach in which we start from an ethical theory and look for the best implementation of it: "takes a specified ethical theory and analyzes its computational requirements to guide the design of algorithms and subsystems capable of implementing that theory'' (Allen et al., 2005, p. 80). In this approach the AMA individual is shaped by its interaction with the community as well as with a set of rules or principles outside itself.

The traditional top-down decision-tree approach to programming involves coding a clearly articulated, domain-specific set of rules that rigidly control a system's behavior. This is the approach that has been followed by Arkin in his work on autonomous weapons, where the rules are taken from the well-established international law of armed conflict (ILOAC) supplemented by context-specific rules of engagement (ROE). A refinement consistent with the top-down approach would involve the addition of an as-needed hedonic algorithm for calculating the balance of pleasure and suffering in specific action settings. That the top-down approach assumes some combination of rule-based, deontological ethical framework with a consequentialist ethical frame is clear and often noted. An obvious advantage of the top-down approach is the ability to build a moral justification of the actions of the AMA. Another advantage is simplicity of implementation. An obvious disadvantage is inflexibility. Another is limitation in scope. One cannot write a rule to cover every contingency. Last but not least, the top down approach is not proper for a social competence.

In the bottom-up approach, the AMA programmers discover the ethical principles by exploring and learning from behavior and actions that are morally praiseworthy. In this approach, the moral sensibility of the individual trumps the rules and restrictions imposed by the programmer. In the bottom up approach the individual is able to discover the social implications of different actions by an internal process.

Finally, there is a hybrid approach. Allen and collaborators reviewed both approaches and endorsed a third, hybrid approach.[11] For all practical purposes, computer scientists may adopt this hybrid approach in which top-down principles and duties imposed from the outside to the AMA are combined with cultivating virtues which are more internal to the agent. The computational model used by Allen et al. is the neural network model provided by connectionism. Our own model, GAMA is a hybrid model and we argue expresses better the social virtues of the GAMA. In this sense, the hybrid model proposed by Allen et al. is in line with our H-4, although our choice for evolutionary programming as expressed by H-3 is different than their approach based on Connectionism Neural Networks (CNN).

The key concept used in the CNN is learning. But can we learn ethical behavior the same way we learn a new language? Some 20th century virtue ethicists (MacIntyre, 1984; Murdoch, 1970) believe that humility is the virtue needed to learn and that the learner approaches the moral domain similarly to the way we learn a new domain or a new language by appealing to an external authority (teacher, tradition, religion).

It is our assumption that "evolutionary" or "genetic" algorithms are able to develop ethical competence not by coding in rules or principles from the beginning, but by making the agent develop and improve a set of ethical habits. Our implementation applies a different principle: "the authority" used by Murdoch for example is as an input: a set of initial behaviors, or a populations of AMAs which are taken to be acceptable or "right" by a majority of human moral agents. From these behaviors the algorithm will produce subsequent generations of behaviors which will combine in a set of internal virtues of the AMA. Also the authority is used as a termination condition for the training period.

Socially speaking, humans are the initial condition of the genetic algorithm of the GAMA and its termination condition. So the GAMA autonomy is both restricted and special: the fitness function, the optimality and the termination are chosen by the human trainer. But the result itr itself, as a set of virtues developed during the sequence of generation is internal to the GAMA.

In the virtue ethics model adopted here, the intrinsic rightness or wrongness of this or that specific action are not crucial, but, instead, the question of the virtuous or vicious character of those settled dispositions or habits of action that are productive of individual actions. It is argued that this habit-based approach to ethics more accurately reflects the nature of mature, human moral competence - we act out of developed habit not in consequence of the execution of a practical syllogism - and thus holds greater promise of creating in artificial moral agents the same kind of moral competence as seen in human agents.

---

11 See Ch 8 in (Allen & Wallach, 2009).

Our model challenge the idea that moral virtues can be taught or learned by following a set of rules: the virtue ethics model adopted here is more by development than by acquirement. The individual AMA develops and achieves a set of virtues which only originate in others' behavior.

The GAMA project in short: two implementations

In the spirit of H-3 and H-4, we propose here two AMA implementations. In both cases the virtue ethics model is implemented fundamentally as a *search for optimality* procedure; arguably, genetic algorithms are the most efficient form of computation. The core idea is a two-step process: first, the algorithm we build a system of virtues from a given set of known and accepted individuals (they can be a configuration behavior or a set of AMAs as neural networks). Second, for a new, previously unknown ethical decision, we use the virtue ethics model to decide based on the system of virtues which action is right. The AMA starts from a set of known behavior socially acceptable to humans. But from this basis, the AMA builds new, previously unknown behaviors that develops during multiple generations. A human trainer choose the termination moment.

The suggestion here is that aspects of ethical behaviors can be recombined and if necessary eliminated in building "the habit" of an AMA. We argue and we illustrate with concrete examples that the advantages of evolutionary programming trump the difficulties mentioned above, especially when we deal with the process of optimizing and refining a set of moral behaviors.

In these implementations, the output of the algorithm is a set of virtual virtues, either of a sole GAMA or of a population of AMAs. Classes of moral behaviors generate classes of virtues which correspond in evolutionary algorithms to species of individuals.

GA component:	GAMA1: Population of AMAs	GAMA2: Population of behaviors
Individuals	AMAs, each having a neural network configuration solving a simple problem	Configuration behavior of one AMA on simple examples with Y/N answers ("trolley like")
Initial population	A set of AMAs created by rule-following	Behaviors for a given problem
Environment	fixed	fixed
Generation +1	New AMAs with new neural networks	New behavior, for a given problem
Optimization/fitness	The result to a given problem P	The result to a given problem P
Termination of decision	Pareto front	Pareto front

[12] Be it in Plato's *Meno* or Aristotle's *Nicomachean Ethics*, this is a well-known problem.

Termination of training	"age of majority" "graduation of the population of AMA	The right balance of virtues for P

A final note is in order here. At this stage, the GAMA project is still a highly idealized model. The environment, i.e. the moral problem P, is fixed and the AMA is trying to find the optimal answer to it. It does not assume a changing environment, but the authors considered also the Evolutionary Adaptation alternative in which the problem becomes more complex during the training period (Jong, 2006, sec. 5.6).

Conclusion

We can think of H-3 and H-4 as ways of replacing the language of rule-following in ethical programing (as proposed by Arkin, Anderson&Anderson, etc.) with something along the lines of optimization and search procedure. The process itself is close to learning, but can be better described as self-building, or better self-discovering, of an increasingly optimal set of values from previous data. In this sense the suggestion to link genetic programming to ethical programming when one adopts the virtue ethics model comes naturally. The social aspect of the GAMA is its constant interaction with a human trainer which decides which generation of AMA is suitable enough. The trainer of GAMA brings in a system of social values and criteria which is not present in the initial population of behaviors. As a project in artificial life and machine ethics, the GAMA proposal brings in two new elements: the virtue ethics model for artificial moral agents and the evolutionary computation. The proposal, still at a prospective stage, may be very difficult to implement. The main advantage is that it may answer some foundational issues in machine ethics the GAMA machine is able to build a developmental set of "virtual virtues". Social abilities of GAMAs are also consequences of the interaction between the GAMA and the trainer. Evolutionary computation, together with the virtue ethics model brings social robotics

References

Affenzeller, M. (2009). *Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications*. Boca Raton, Fla: CRC Press.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, *7*(3), 149–155. doi:10.1007/s10676-006-0004-4

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, *12*(3), 251–261. doi:10.1080/09528130050111428

Allen, C., & Wallach, W. (2009). *Moral machines: teaching robots right from wrong*. Oxford; New York: Oxford University Press.

Anderson, M., & Anderson, S. L. (Eds.). (2011). *Machine Ethics*. Cambridge University Press.

Anderson, M., Anderson, S. L., & Gips, J. (Eds.). (2011). Towards the Ethical Robot. In *Machine Ethics*. Cambridge University Press.

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*. CRC Press.

Arkin, R. C. (1998). *Behavior-based robotics*. Cambridge, Mass.: MIT Press.

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, *29*(1), 94–103.

Casebeer, W. D. (2005). *Natural ethical facts: evolution, connectionism, and moral cognition*. Cambridge, Mass.; London: MIT.

Chalmers, D. (2010). The Singularity: A philosophical analysis. *Journal of Consciousness Studies*, *17*, *9*(10), 7–65.

Churchland, P. M. (1995). *The engine of reason, the seat of the soul : a philosophical journey into the brain*. Cambridge, Mass.: MIT Press.

Danielson, P. (1992). *Artificial morality virtuous robots for virtual games*. London; New York: Routledge. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=68930

Economist, T. (2014, January 18). The onrushing wave. *The Economist*. Retrieved from http://www.economist.com/news/briefing/21594264-previous-technological-innovation-has-always-delivered-more-long-run-employment-not-less?frsc=dg%7Ca

Fogel, L., Owens, A., & Walsh, M. (1966). *Artificial Intelligence through Simulated Evolution*. John Wiley.

Ford, K. M., Glymour, C. N., & Hayes, P. J. (1995). *Android epistemology*. Menlo Park; Cambridge, Mass.: AAAI Press ; MIT Press.

Gips, J. (1995). Towards the ethical robot. In K. M. Ford, C. N. Glymour, & P. J. Hayes (Eds.), *Android epistemology*. Menlo Park; Cambridge, Mass.: AAAI Press ; MIT Press.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (second edition Bradford Books, 1992.). University of Michigan Press.

Hursthouse, R. (1999). *On virtue ethics*. Oxford; New York: Oxford University Press.

Jong, K. A. de D. (2006). *Evolutionary Computation* (1st ed.). Cambridge MA: MIT Press: A Bradford Book.

Koza, J. R., III, F. H. B., Andre, D., & Keane, M. A. (Eds.). (1999). *Genetic Programming III: Darwinian Invention and Problem Solving* (1st ed.). Morgan Kaufmann.

Koza, J. R., Keane, M., Streeter, M., Mydlowec, W., Yu, J., & Lanza, G. (Eds.). (2003). *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. Norwell, Mass: Kluwer Academic Publishers.

Kurzweil, R. (2006). *The singularity is near: when humans transcend biology*. New York: Penguin.

MacIntyre, A. C. (1984). *After virtue: a study in moral theory*. Notre Dame, Ind.: University of Notre Dame Press.

Matthijs Pontier. (2014). *Moral Coppélia: Affective moral reasoning with twofold autonomy and...* MEMCA14, at AISB50. Retrieved from http://www.slideshare net/Matthijs85/moral-copplia-affective-moral-reasoning-with-twofold-autonomy-and-a-touch-of-personality-presentation-at-memca14-symposium-at-aisb50

Murdoch, I. (1970). *The sovereignty of good*. London; New York: Routledge.

Olariu, S., & Zomaya, A. Y. (Eds.). (2006). *Handbook of Bioinspired Algorithms and Applications* (1st ed.). Chapman and Hall/CRC.

Stone, R. (2013). Scientists Campaign Against Killer Robots. *Science*, *342*(6165), 1428 – 1429. doi:10.1126/science.342.6165.1428

Tomassini, M. (1995). A Survey of Genetic Algorithms. *Annual Reviews of Computational Physics*, *3*, 87–118.

Turing, A. (1996). Intelligent Machinery, A Heretical Theory. *Philosophia Mathematica*, *4*(3), 256–260.

Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding and selection in evolution. In *Proc of the 6th International Congress of Genetics* (Vol. 1, pp. 356–366).